HuITBERT: A Vietnamese Pretrained Encoder for Long-Context Understanding

Anonymous ACL submission

Abstract

Vietnamese pretrained encoders remain limited by short context windows and outdated design choices. We introduce HuITBERT, a Vietnamese monolingual encoder pretrained from scratch under the ModernBERT design with native support for contexts up to 4,096 tokens, using a 322 GB corpus from FineWeb-2 and FineWiki and a Vietnamese-specific 32K BPE tokenizer (Penedo et al., 2025). On a unified benchmark of nine Vietnamese NLU tasks from eight datasets, HuITBERT$_{base}$ achieves the best average score while using only 134M parameters, slightly surpassing larger encoders such as mmBERT$_{base}$, XLM-RoBERTa$_{base}$, and PhoBERT$_{large}$ (Marone et al., 2025; Conneau et al., 2020; Nguyen and Nguyen, 2020). We additionally conduct analyses to demonstrate its effectiveness on other tasks. These results establish HuITBERT as a strong and efficient Vietnamese encoder for Vietnamese language understanding. This initiative aims to advance research and improve pretrained language models for low-resource languages [1].

1 Introduction

Pretrained language models are the foundation of modern natural language understanding (NLU), and encoder-only architectures remain especially strong for classification, retrieval, and span prediction (Devlin et al., 2019; Liu et al., 2019). In Vietnamese NLP, this line of work has produced strong monolingual encoders such as PhoBERT (Nguyen and Nguyen, 2020), ViDeBERTa (Tran et al., 2023), ViSoBERT, and ViHealthBERT (Nguyen et al., 2023; Minh et al., 2022), while multilingual encoders such as XLM-RoBERTa remain competitive because of their scale and transferability (Conneau et al., 2020). However, most Vietnamese encoders still rely on short context windows and older pretraining designs.

Recent work shows that encoder models remain highly competitive when modernized. ModernBERT combines long-context support with improved efficiency through architectural and systems-level changes such as Rotary Positional Embeddings, alternating local-global attention, and efficient attention kernels (Warner et al., 2025; Su et al., 2023; Dao et al., 2022). Similar trends also appear in newer language-specific and multilingual encoders (Breton et al., 2025; Sugiura et al., 2025; Marone et al., 2025). However, this design space has not yet been systematically explored for a Vietnamese monolingual encoder pretrained from scratch.

This gap matters because Vietnamese remains under-served compared with high-resource languages, and many downstream benchmarks now require stronger contextual representations for inference, question answering, long-document evidence aggregation, and retrieval-oriented semantic matching (Do et al., 2024; Tran et al., 2024; Ngo et al., 2024; Hoa et al., 2025). A Vietnamese encoder that combines monolingual specialization with modern long-context efficiency would therefore be practically useful.

In this paper, we introduce HuITBERT, to the best of our knowledge the first Vietnamese monolingual encoder pretrained from scratch under the ModernBERT design with native support for contexts up to 4,096 tokens. We pretrain HuITBERT on a 322 GB Vietnamese corpus derived from FineWeb-2 and FineWiki and train a Vietnamese-specific 32K BPE tokenizer (Penedo et al., 2025; Rust et al., 2021; Gallé, 2019). We study two model sizes, Tiny and Base, using a two-stage recipe that first learns general bidirectional representations at shorter context and then extends the context window to 4,096 tokens.

We evaluate HuITBERT on a unified Vietnamese NLU benchmark covering nine tasks from eight datasets across classification, sentence-pair reason-

[1] The supplementary material will be made publicly available upon acceptance

ing, machine reading comprehension, and structured prediction. HuITBERT$_{base}$ achieves the best average performance, slightly outperforming mmBERT$_{base}$ (Marone et al., 2025) while using less than half the parameters, and surpassing strong Vietnamese and multilingual baselines such as PhoBERT$_{large}$ (Nguyen and Nguyen, 2020) and XLM-RoBERTa$_{base}$ (Conneau et al., 2020).

We further evaluate HuITBERT on dense retrieval over Vietnamese benchmarks (ViQuAD2.0 (Kiet et al., 2022) and Zalo-Legal[2]), where it substantially outperforms several Vietnamese embedding baselines, reaches near parity with a specialized Vietnamese bi-encoder on UIT-ViQuAD2.0, and remains below that model on Zalo-Legal. Finally, we show that longer context windows improve performance on VlogQA and ViFactCheck while maintaining competitive inference throughput. Our contributions are as follows:

- We introduce HuITBERT, to the best of our knowledge the first Vietnamese monolingual encoder pretrained from scratch under the ModernBERT design with native support for contexts up to 4,096 tokens.
- We build a 322 GB Vietnamese pretraining corpus and train a Vietnamese-specific 32K BPE tokenizer for large-scale monolingual encoder pretraining.
- We show that HuITBERT$_{base}$ achieves the best average performance on a unified benchmark of nine Vietnamese NLU tasks from eight datasets.
- We provide additional analyses showing its effectiveness on other tasks and the trade-off between short- and long-context modeling.

2 Related Work

2.1 Foundational Pre-trained Language Models

Encoder-only Transformers, introduced with BERT (Devlin et al., 2019), constitute the cornerstone of Natural Language Understanding (NLU). Variants such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) further optimized pre-training objectives to maximize bidirectional representation learning. Despite the dominance of autoregressive decoders, encoders maintain superior inference efficiency and consistently outperform larger decoder-only models on retrieval and classification benchmarks (Weller et al.; Nielsen et al., 2025).

To address the suboptimal model design of foundational encoders (Warner et al., 2025), recent research has integrated architectural advancements from large language models (LLMs) to enhance context capacity and inference efficiency. ModernBERT (Warner et al., 2025) and NeoBERT (Breton et al., 2025) adopt Rotary Positional Embeddings (RoPE) to extend context from 512 to 8,192 tokens while maintaining efficiency through alternating global-local attention. Gated activations (GeGLU/SwiGLU) and unpadding techniques yield 2× faster inference versus DeBERTaV3 (He et al., 2020).

For language-specific optimization, Chinese ModernBERT (Zhao et al., 2025) employs a 32k BPE vocabulary with dynamic masking (30% → 15%) and Whole-Word Masking for improved tokenization. ModernGBERT (Ehrmanntraut et al., 2025) and NeoDictaBERT (Shmidman et al., 2025) show that training modern encoder architectures from scratch can outperform simpler adaptations from older backbones. Similarly, `llm-jp-modernbert` (Sugiura et al., 2025) applies the architecture to Japanese and studies long-context adaptation under native 8,192-token support. At multilingual scale, mmBERT (Marone et al., 2025) uses Cascading Annealed Language Learning to support 1,800+ languages, while EuroBERT (Boizard et al., 2025) studies masking-ratio trade-offs during pretraining. These works collectively suggest that architectural modernization, tokenizer design, and long-context adaptation are complementary rather than interchangeable factors.

2.2 Vietnamese Research on Pre-trained Language Models

Table 1 provides an overview of representative multilingual and Vietnamese pretrained language models, comparing their architectures, training corpora, and tokenization strategies.

Early research in Vietnamese NLP primarily relied on massive multilingual encoders. While models like mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) offered zero-shot transferability, they often lacked deep alignment with Vietnamese linguistic nuances.

To address these limitations, the field shifted toward dedicated monolingual architectures with varying tokenization strategies and data scales.

[2] https://challenge.zalo.ai/

PhoBERT (Nguyen and Nguyen, 2020), adapting RoBERTa with explicit word segmentation, was pre-trained on a 20GB corpus of news and Wikipedia data. Alternative approaches included ViBERT (Bui et al., 2020), trained on 10GB of news text, and vELECTRA (Bui et al., 2020), which utilized a generator-discriminator objective on 60GB of data to enhance training efficiency. Pushing the scale further, ViDeBERTa (Tran et al., 2023) leveraged 138GB of data from CC100 and a disentangled attention mechanism or CafeBERT (Do et al., 2024) continued pretraining base on XLM-Roberta for improve on Vietnamese task. Concurrently, domain-adaptive models have emerged to handle specialized vocabularies, including ViSoBERT (Nguyen et al., 2023) for social media, ViHealthBERT (Minh et al., 2022) for healthcare.

Despite these advancements, existing Vietnamese encoders remain largely constrained by pre-LLM architectural designs, characterized by shorter context windows and pretraining recipes that do not explicitly target long-context adaptation. While multilingual modern encoders such as SEA-LION-ModernBERT provide a relevant regional baseline, they do not replace the need for a Vietnamese monolingual encoder trained from scratch with Vietnamese-specific tokenization and pretraining data. This gap motivates the development of a Vietnamese-specific encoder based on the ModernBERT design to support efficient, long-context retrieval and reasoning.

3 Methodology

3.1 Training Corpus

Our pretraining corpus combines two complementary, high-quality sources. **FineWeb-2** (Penedo et al., 2025), derived from nearly 100 Common Crawl snapshots (mid-2013 to April 2024), provides large-scale and temporally diverse web text; we use approximately 320 GB of cleaned Vietnamese data. To improve factual coverage, we additionally include **FineWiki** [3], a structurally normalized encyclopedic resource with approximately 1.28M pages (approximately 2.1 GB) that addresses limitations of standard Wikipedia XML dumps. Together, these sources balance broad linguistic coverage with dense factual content. For the two-stage schedule, we materialize two stage-specific sampled corpora: Stage 1 uses 1,024-token chunks and contains 65.97B training tokens, while Stage 2 uses 4,096-token chunks and contains 22.39B training tokens. Appendix B reports the corresponding sequence-length distributions for both stages. We retain the dataset-level filtering and deduplication provided by the underlying sources, but we do not claim a complete audit of licensing, personally identifiable information, or toxic content; we therefore treat data curation as a practical but incomplete safeguard rather than a solved problem.

[3] https://huggingface.co/datasets/HuggingFaceFW/finewiki

3.2 Tokenization Strategy

Following prior work emphasizing language-specific tokenization (Rust et al., 2021), we train a Byte-Pair Encoding (BPE) tokenizer (Gallé, 2019) on the pretraining corpus. Unlike prior Vietnamese encoders that heavily rely on explicit word pre-segmentation (Nguyen and Nguyen, 2020; Tran et al., 2023), we apply BPE directly to the raw, unsegmented text. This end-to-end approach simplifies the preprocessing pipeline and prevents cascading errors from external word segmenters, making the tokenizer more robust to the noisy nature of web-scale data. We set the vocabulary size to 32K subword units and round it to be divisible by 64 for computational efficiency. The tokenizer follows the standard BERT input format and special tokens (e.g., `[CLS]` and `[SEP]`) (Devlin et al., 2019). Inspired by GPT-NeoX-20B (Black et al., 2022), we additionally model repeated whitespace tokens to improve robustness to noisy, web-scale text.

Consistent with findings from SindBERT (Scheible-Schmitt and Schweter, 2025) and LLäMmlein (Pfister et al., 2025), we train the tokenizer on a stable subset of the corpus rather than the full dataset to reduce training cost. As shown in Appendix A, tokenizer quality (fertility and percentage of continued words; pcw) (Rust et al., 2021) saturates early with increasing tokenizer training data size and yields competitive segmentation efficiency relative to existing Vietnamese tokenizers.

3.3 Architecture Overview

HuITBERT follows the ModernBERT (Warner et al., 2025) design, incorporating Rotary Positional Embedding (RoPE) (Su et al., 2023), Local–Global Alternating Attention (Team, 2024), and FlashAttention (Dao et al., 2022) for efficient long-context modeling. We pretrain with a maximum context length of 4,096 tokens and consider two

Table 1: Comparison of Vietnamese pretrained language models. *MSL denotes the maximum sequence length supported by the model.

Model	Backbone	Params	MSL	Training Data	Data Source	Vocab	Tokenizer	Pre-seg.
				Multilingual				
mBERT	BERT-Base	177M	512	_	Wikipedia	119K	WordPiece	No
XLM-R$_{base}$	RoBERTa	270M	512	138 GB	Common Crawl	250K	SentencePiece	No
				Vietnamese				
PhoBERT$_{base}$	RoBERTa	135M	512	20 GB	Wikipedia, News	64K	BPE	Yes
PhoBERT$_{large}$	RoBERTa	370M	512	20 GB	Wikipedia, News	64K	BPE	Yes
ViBERT	BERT	115M	512	10 GB	News	38K	WordPiece	No
ViDeBERTa$_{xsmall}$	DeBERTaV3	70M	512	138 GB	CC-100	128K	SentencePiece	Yes
ViDeBERTa$_{base}$	DeBERTaV3	184M	512	138 GB	CC-100	128K	SentencePiece	Yes
ViSoBERT	XLM-R	98M	512	1 GB	Social Media	15K	SentencePiece	No
HuITBERT$_{tiny/base}$	ModernBERT	32M/134M	4,096	322 GB	FineWeb, FineWiki	32K	BPE	No

sizes (Tiny and Base); the complete architectural specification is reported in Appendix B (Table 7). For initialization, Tiny uses the Megatron-LM scheme (Shoeybi et al., 2020), while Base Stage 1 is initialized from the Stage 1 Tiny checkpoint using weight tiling from the middle; Base Stage 2 then continues from the resulting Base Stage 1 checkpoint.

3.4 Training Configuration

We pretrain both variants with a two-stage recipe: (i) an initial pretraining stage at shorter context and (ii) a context-extension stage that continues training at 4,096 tokens with updated RoPE settings. Stage 1 uses the 1,024-token corpus, Stage 2 uses the 4,096-token corpus, and each stage is trained for four epochs over its respective training split. Unless stated otherwise, we use standard masked language modeling (MLM) with dynamic masking.

To avoid repeating fine-grained hyperparameters in the main text, we report the full stage-wise pretraining configuration in Appendix B (Table 9).

We set the training masking probability to 0.30 and the evaluation masking probability to 0.15, and we exclude padding tokens from the loss computation. Sequence packing is enabled for training to improve throughput but disabled for evaluation. Packed examples preserve document boundaries through attention masking, so tokens from different source documents do not attend to one another during MLM pretraining. We use the same RoPE θ values reported in Appendix B; Stage 2 extends the usable context window by continuing pretraining at longer sequence length rather than by changing the backbone architecture.

System and reproducibility. All pretraining experiments use automatic mixed precision (bfloat16) and a fixed random seed of 17 to ensure stable and reproducible optimization. Training is performed on RTX A6000 GPUs (48GB): 4× for the Tiny model and 8× for the Base model. We report implementation-level throughput statistics for this multi-GPU setup in Appendix C; these numbers should be interpreted as measured system throughput within our training stack rather than as hardware-independent efficiency claims. Additional reproduction details are provided in Appendix C.

4 Experiments

4.1 Downstream tasks

Recent efforts have been devoted to constructing evaluation datasets for Vietnamese natural language understanding (NLU), including composite benchmarks such as VLUE (Vietnamese Language Understanding Evaluation) (Do et al., 2024) and ViGLUE (Vietnamese General Language Understanding Benchmark) (Tran et al., 2024). Despite their contributions, these benchmarks remain limited in task diversity and are affected by translationese, as a substantial portion of the data is derived from translated sources rather than originally authored Vietnamese text. To address these limitations, we assemble a downstream evaluation collection consisting of eight datasets and nine evaluation tasks organized into four principal task categories: single sentence, sentence pair, machine reading comprehension (MRC), and structured prediction. The nine tasks arise because VSFC is evaluated on both sentiment classification and topic classifica-

tion. This collection is designed to provide a more comprehensive assessment of Vietnamese NLU capabilities across heterogeneous task formulations and data sources, thereby enabling a more fine-grained analysis of model performance beyond existing benchmarks. Table 2 reports basic statistics for all datasets used in our downstream evaluation.

Table 2: Statistics of Vietnamese NLP Benchmark Datasets

Dataset	Domains	Train	Dev	Test
VSMEC	Social Media	5,548	686	693
VSFC	Education	11,426	1,583	1,900
VToC	News	7,293	-	1,831
ViHSD	Social Media	24,048	2,672	6,680
ViQuAD2.0	Wikipedia	22,765	5,692	3,712
ViNLI	News	24,376	3,009	2,991
PhoNER	Healthcare	5,027	2,000	3,000
VLSP-2013	General	27,000	770	2,120

The main benchmark covers four task categories: single-sentence classification (VToC, VSFC Sentiment, VSFC Topic, VSMEC, and ViHSD), sentence-pair inference (ViNLI), machine reading comprehension (ViQuAD2.0), and structured prediction (PhoNER and VLSP-2013 POS tagging). We move the dataset-level descriptions of these downstream tasks to Appendix E to keep the main text focused on the evaluation protocol and aggregate results.

4.2 Experimental Setup

For the main benchmark, we fine-tune each pre-trained model on all tasks using a batch size of 32, a maximum sequence length of 512, and 10 training epochs. For each experiment, we run downstream fine-tuning three times with random seeds 10, 42, and 100 to improve the reliability of the results. We use AdamW (Loshchilov and Hutter, 2017) with a linear learning-rate schedule with warmup, selecting the learning rate per task in the range 1×10^{-5}–5×10^{-5}. Early stopping is applied with a patience of 2 epochs based on the development set. All experiments are implemented with the HuggingFace Transformers library (Wolf et al., 2020); full task-specific configurations and shared hyperparameters are provided in Appendix D.

For dense retrieval, we fine-tune bi-encoder models on a contrastive training set built from mMARCO (Bonifacio et al., 2021) and additional legal-domain hard negatives. Specifically, we construct approximately 530K triplets by aligning Vietnamese mMARCO query-passage pairs with the pre-mined hard negatives [4] and further augment training with 37K hard-negative triplets released for the Legal Text Retrieval Zalo 2021 challenge [5].

We follow the Sentence-BERT bi-encoder framework (Reimers and Gurevych, 2019) and optimize the model using MultipleNegativesRankingLoss (van den Oord et al., 2019) with in-batch negatives. Models are trained with a batch size of 64, a learning rate of 5×10^{-5}, for 5 epochs.

Retrieval performance is evaluated using NDCG@10 on ViQuAD2.0 (Kiet et al., 2022) and Zalo-Legal, to assess the transferability of HuIT-BERT representations to semantic retrieval tasks.

4.3 Results on NLU Tasks

Table 3 reports the main benchmark comparison across nine Vietnamese NLP tasks spanning sequence labeling, sentence classification, sentence-pair inference, and machine reading comprehension. Overall, HuITBERT$_{base}$ achieves the best average performance among all compared models, with an AVG score of 81.77. The margin over the strongest baseline is small but consistent: HuITBERT$_{base}$ slightly outperforms mmBERT$_{base}$, and more clearly surpasses PhoBERT$_{large}$ and XLM-RoBERTa$_{base}$. This result is especially meaningful from an efficiency perspective. With only $134M$ parameters, HuITBERT$_{base}$ remains far smaller than several leading baselines, including mmBERT$_{base}$ ($306M$), SEA-LION-ModernBERT ($300M$), XLM-RoBERTa$_{base}$ ($270M$), and PhoBERT$_{large}$ ($370M$). The strongest overall AVG therefore comes from a better accuracy–efficiency trade-off rather than from scaling model size alone.

At the task level, HuITBERT$_{base}$ obtains the best score on ViNLI and VToC, while ranking second on ViHSD and ViQuAD2.0 These gains are distributed across heterogeneous task families, including social media classification, topic classification, sentence-pair reasoning, and extractive question answering, which suggests that the proposed model transfers robustly across diverse Vietnamese language understanding settings. Other baselines remain strongest on specific tasks—for example, mmBERT$_{base}$ on PhoNER and ViQuAD2.0, XLM-RoBERTa$_{base}$ on VLSP POS and VSFC Sentiment, and ViSoBERT on VSMEC, ViHSD, and VSFC

[4] https://huggingface.co/datasets/sentence-transformers/msmarco-hard-negatives

[5] https://huggingface.co/datasets/batmangiaicuuthegioi/zalo-legal-triplets

Table 3: Benchmark results on Vietnamese NLP tasks. The best score in each column is in **bold**, and the second-best score is in *italics*. AVG denotes the average score across all tasks, and results are reported as mean ± standard deviation over multiple runs with different random seeds.

Model	PhoNER	VSMEC	ViHSD	ViNLI	ViQuAD2.0	VLSP POS	VSFC Sent.	VSFC Topic	VToC	AVG
ViSoBERT	97.01	**67.52**	**88.37**	64.17	29.89	93.26	91.48	**88.28**	81.44	77.94 $\pm_{0.49}$
ViDeBERTa$_{xsmall}$	93.22	31.53	82.51	58.31	30.19	89.23	83.81	83.25	28.85	64.54 $\pm_{0.32}$
ViDeBERTa$_{base}$	94.19	22.76	79.09	60.98	29.46	91.20	82.93	81.03	28.83	63.39 $\pm_{0.73}$
PhoBERT$_{base}$	96.88	58.19	85.44	65.61	38.84	92.47	90.90	86.30	82.75	77.49 $\pm_{2.78}$
PhoBERT$_{large}$	97.68	*64.72*	86.60	75.56	43.91	93.65	91.17	87.59	83.48	80.48 $\pm_{0.59}$
ViBERT$_{base}$	96.53	46.26	85.12	67.09	32.60	91.50	89.01	86.02	74.16	74.25 $\pm_{0.28}$
BERT$_{base}$	97.60	53.11	85.45	72.16	40.52	94.03	90.90	86.77	81.87	78.05 $\pm_{0.23}$
XLM-RoBERTa$_{base}$	*97.69*	61.23	86.64	75.83	38.73	**94.47**	**91.70**	87.49	83.43	79.69 $\pm_{0.17}$
SEA-LION-ModernBERT [9]	96.40	52.11	85.58	75.24	48.03	92.88	89.28	87.37	81.48	78.71 $\pm_{0.15}$
mmBERT$_{base}$	**97.73**	60.80	86.16	*80.47*	**51.87**	*94.12*	*91.53*	*88.09*	*83.97*	*81.64* $\pm_{0.09}$
HuITBERT$_{tiny}$	96.94	57.40	86.45	68.91	34.67	93.47	89.61	86.42	81.09	77.22 $\pm_{0.17}$
HuITBERT$_{base}$	97.54	63.95	*87.68*	**80.74**	*49.07*	93.92	91.30	87.44	**84.32**	**81.77** $\pm_{0.04}$

Topic—but none matches the same level of consistency across the full benchmark. Compared with similarly sized monolingual encoders such as PhoBERT$_{base}$ and ViBERT$_{base}$, HuITBERT$_{base}$ shows a clear improvement in overall generalization while remaining competitive with substantially larger multilingual models.

The comparison between the two HuITBERT variants shows that the proposed recipe also scales well to compact models. Although HuITBERT$_{tiny}$ has only 32M parameters, it still reaches an AVG of 77.22, outperforming larger baselines such as ViDeBERTa$_{xsmall}$, ViDeBERTa$_{base}$, and ViBERT$_{base}$, while remaining close to PhoBERT$_{base}$ and ViSoBERT. This result indicates that the architectural and pretraining choices remain effective even in a lightweight setting. Taken together, these findings support the view that HuITBERT is a strong general-purpose Vietnamese encoder family, with HuITBERT$_{base}$ offering the best overall balance between task performance and parameter efficiency.

4.4 Results on Information Retrieval Tasks

We further evaluate whether HuITBERT transfers beyond standard NLU benchmarks to dense retrieval. To this end, we fine-tune the models on the combined Vietnamese mMARCO and Zalo-legal hard-negative training set using MultipleNegativesRankingLoss and report NDCG@10 on ViQuAD2.0 and Zalo-Legal.

Table 4 shows that HuITBERT$_{base}$ transfers strongly to dense retrieval under this practical evaluation setting. It outperforms the public Vietnamese-Embedding, Sup-SimCSE PhoBERT, and Vietnamese-SBERT checkpoints on both benchmarks, improving over Vietnamese-SBERT by 9.13 points on UIT-ViQuAD2.0 and 18.55 points on Zalo-Legal. The gains are even larger relative to the supervised SimCSE PhoBERT baseline, especially on the legal-domain benchmark.

Table 4: NDCG@10 on Vietnamese information retrieval benchmarks. The best score in each column is in **bold**, and the second-best score is in *italics*.

Model	ViQuAD2.0	Zalo-Legal
Vietnamese-Embedding [10]	57.31	36.35
Sup-SimCSE PhoBERT [11]	55.53	30.64
Vietnamese-SBERT [12]	56.66	37.36
HuITBERT$_{tiny}$	58.38	31.49
HuITBERT$_{base}$	*65.79*	*55.91*
Vietnamese Bi-Encoder [13]	**65.99**	**61.37**

At the same time, the results show a meaningful domain gap. On UIT-ViQuAD2.0, HuITBERT$_{base}$ is nearly tied with the specialized Vietnamese Bi-Encoder baseline (Duc et al., 2024) (65.79 vs. 65.99), indicating that the pretrained encoder transfers well to retrieval after contrastive fine-tuning. On Zalo-Legal, however, HuITBERT$_{base}$ remains below the same specialized bi-encoder (55.91 vs. 61.37), suggesting that retrieval in specialized domains remains sensitive to retrieval-oriented pretraining or domain adaptation. The much weaker results of HuITBERT$_{tiny}$ further indicate that dense retrieval is more demanding than the aggregate NLU benchmark and benefits substantially from the larger base model.

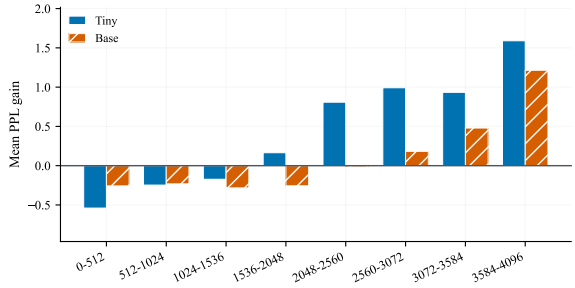

Figure 1: Mean PPPL gain of Stage 2 over Stage 1 across sequence-length bins.

4.5 Analysis

4.5.1 Long-Context pseudo-perplexity

Following Sugiura et al. (2025); Breton et al. (2025), for a tokenized sequence $x = (x_1, \ldots, x_L)$, let T denote the set of special tokens (e.g., [CLS], [SEP], [PAD], [UNK]), and define the set of content positions as

$$S = \{i \in \{1, \ldots, L\} : x_i \notin T\}. \tag{1}$$

Let $x^{(i)}$ denote the sequence obtained by replacing x_i with [MASK]. The loss at position i is

$$\ell_i = -\log p_\theta(x_i \mid x^{(i)}), \qquad i \in S. \tag{2}$$

We define pseudo-perplexity (PPPL) as

$$\mathrm{PPPL}(x) = \exp\left(\frac{1}{|S|}\sum_{i \in S} \ell_i\right). \tag{3}$$

For long sequences, we compute $\{\ell_i\}_{i \in S}$ efficiently by evaluating multiple masked positions in parallel within each forward pass.

Figure 1 shows a clear length-dependent effect of Stage 2. For sequences up to 1536 tokens, Stage 2 increases pseudo-perplexity for both HuITBERT$_{tiny}$ and HuITBERT$_{base}$, indicating no short-context benefit. Beyond this range, it becomes beneficial: HuITBERT$_{tiny}$ improves from 1536–2048 tokens onward, whereas HuITBERT Base improves later, with the largest gains beyond 2048 tokens. This delayed crossover for HuITBERT$_{base}$ may reflect its stronger short-context initialization from the tiled Stage 1 Tiny checkpoint. Overall, Stage 2 mainly improves long-context modeling rather than uniformly improving all sequence lengths.

4.5.2 Long-Context Understanding

We evaluate long-context understanding on two Vietnamese benchmarks: VlogQA (Ngo et al., 2024), which tests question answering over long, noisy spoken transcripts, and ViFactCheck (Hoa et al., 2025), which evaluates claim verification from document context. We compare HuITBERT under 1024- and 4096-token settings and report length-stratified results. Each checkpoint is fine-tuned at its corresponding maximum length, and inputs longer than the active limit are truncated. Appendix D lists the task-specific configurations.

VlogQA tests robustness on long spoken contexts with conversational filler and transcription noise. Figure 2 shows EM across input-length buckets.

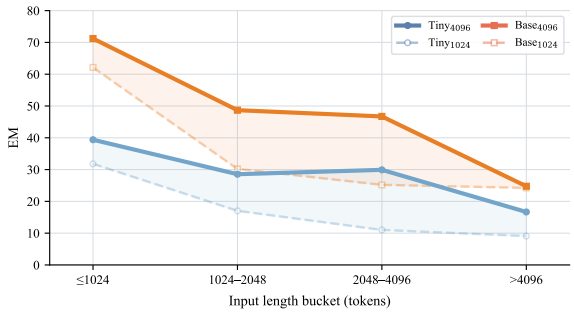

Figure 2: VlogQA EM by context length.

Increasing the context window from 1024 to 4096 tokens consistently improves VlogQA EM, with the largest gains in the 2048–4096 token range: +18.90 for Tiny and +21.52 for Base. This suggests that longer context helps when answer evidence is spread across extended transcripts.

ViFactCheck evaluates claim verification from textual evidence. Figure 3 reports F1 for short and long contexts.

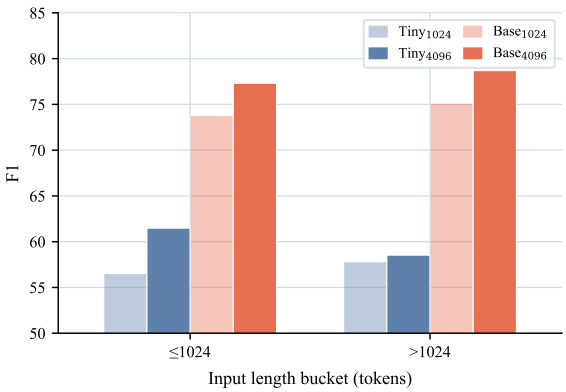

Figure 3: ViFactCheck F1 by context length.

Longer context also improves factual verification. The 4096-token models outperform their 1024-token counterparts on both short and long

inputs, with gains of +3.87 F1 for Tiny and +3.54 for Base. This suggests better retention and integration of supporting evidence.

Overall, larger context windows improve both answer extraction and evidence-based reasoning on these Vietnamese benchmarks. However, as the pseudo-perplexity analysis above suggests, the same Stage 2 recipe does not uniformly improve shorter contexts.

4.5.3 Inference Efficiency

Table 5 reports implementation-level throughput in seconds per million tokens (lower is better) measured within our inference setup on a single NVIDIA RTX A6000 GPU. We evaluate both fixed-length inputs and variable-length inputs, where short lengths are sampled from a truncated normal distribution $\mathcal{N}(256, 64^2)$ clipped to $[16, 512]$ and long lengths are sampled from $\mathcal{N}(2048, 768^2)$ clipped to $[1024, 4096]$, with padding to the corresponding maximum length.

Table 5: Model throughput in seconds per million tokens.

Model	Short		Long	
	Fixed	Variable	Fixed	Variable
ViSoBERT	3.1716	6.9043	-	-
PhoBERT$_{base}$	3.1338	3.5741	-	-
PhoBERT$_{large}$	9.3824	10.6661	-	-
XLM-RoBERTa$_{base}$	3.1286	6.9365	-	-
XLM-RoBERTa$_{large}$	9.6392	20.4278	-	-
mmBERT$_{base}$	4.4282	4.4080	5.1777	4.8096
SEA-LION-300M	4.3954	4.3922	5.1421	4.7750
HuITBERT$_{tiny}$	1.0951	1.1106	1.3013	1.2267
HuITBERT$_{base}$	4.3300	4.3966	5.1376	4.7619

HuITBERT$_{tiny}$ consistently achieves the best throughput across all settings, while HuITBERT$_{base}$ remains competitive with other base-sized models. The small gap between fixed-length and variable-length throughput demonstrates that the models maintain stable efficiency under variable-length inputs. Moreover, the slight improvement in the long variable-length setting suggests that unpadding helps reduce unnecessary computation on padded tokens. Because these measurements depend on the software stack and batching strategy, we present them as implementation-level comparisons rather than universal hardware-independent efficiency claims.

5 Conclusion & Future Work

We introduced HuITBERT, a Vietnamese monolingual encoder family based on ModernBERT and pretrained from scratch with native support for sequences up to 4,096 tokens. Combining a Vietnamese-specific 32K BPE tokenizer, a large curated corpus, and a two-stage context-extension schedule, HuITBERT$_{base}$ achieves the best average performance across nine tasks from eight Vietnamese datasets while using fewer parameters than several strong multilingual and monolingual baselines. It also transfers effectively to dense retrieval, outperforming several common Vietnamese embedding baselines, reaching near parity with a specialized Vietnamese bi-encoder on UIT-ViQuAD2.0, and remaining below that model on Zalo-Legal. Our analysis on VlogQA and ViFactCheck further shows that longer context windows improve long-document question answering and fact verification, although these gains are not uniform across sequence lengths and come with some short-context trade-offs. Future work should therefore explore better disentanglement of tokenizer, architecture, and training curriculum, as well as improved context-extension strategies that more effectively balance short- and long-context performance.

Future work will extend HuITBERT beyond its current general-domain ModernBERT-style long-context encoder. In particular, future versions can be adapted to high-value domains such as biomedical, clinical, legal, and educational text through continued domain-adaptive pretraining (Gururangan et al., 2020; Sounack et al., 2025). HuITBERT can also be developed into Vietnamese embedding and reranking models for retrieval-augmented systems, following recent progress in multilingual long-context retrieval and late-interaction modeling (Zhang et al., 2024; Chaffin and Sourty, 2025). Finally, HuITBERT may serve as a backbone for trustworthiness-aware RAG pipelines, including encoder-based hallucination detection (Kovács and Recski, 2025).

Limitations

HuITBERT improves long-context Vietnamese language understanding, but the current study has several limitations. First, the gains on the main benchmark are modest in absolute terms, even though they are consistent on average, so the paper should be interpreted as evidence for a better accuracy–efficiency trade-off rather than a large across-the-board improvement on every task. Second, our experiments do not fully disentangle the contributions of architecture, tokenizer, corpus composition, and the two-stage context-extension schedule; targeted ablations remain important future work. Third, the pseudo-perplexity analysis indicates that Stage 2 primarily benefits longer sequences and may weaken short-context modeling, suggesting that mixed-length curricula or alternative context-extension strategies may further improve the short/long trade-off. Fourth, our retrieval evaluation covers only two benchmarks and does not isolate the effect of hard-negative mining, domain adaptation, or alternative retrieval objectives, so stronger conclusions about Vietnamese retrieval should be deferred until broader evaluation. Finally, like other web-scale pretrained models, HuITBERT inherits potential noise, bias, and harmful content from web data despite corpus curation. In particular, we do not claim a comprehensive pipeline for detecting or removing personally identifiable information (PII), so a more complete audit of privacy risks, licensing, and downstream harms is still needed.

Acknowledgments

References

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, Joao Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno Miguel Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages. *arXiv.org*.

Luiz Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset. *CoRR*, abs/2108.13897.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, and Sarath Chandar. 2025. Neobert: A next-generation bert. *arXiv.org*.

The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 2020. Improving sequence tagging for Vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 13–20, Hanoi, Vietnam. Association for Computational Linguistics.

Antoine Chaffin and Raphaël Sourty. 2025. Pylate: Flexible training and retrieval for late interaction models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Phong Nguyen-Thuan Do, Son Quoc Tran, Phu Gia Hoang, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. VLUE: A new benchmark and multi-task knowledge transfer learning for Vietnamese natural language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 211–222, Mexico City, Mexico. Association for Computational Linguistics.

Nguyen Quang Duc, Le Hai Son, Nguyen Duc Nhan, Nguyen Dich Nhat Minh, Le Thanh Huong, and Dinh Viet Sang. 2024. Towards comprehensive vietnamese retrieval-augmented generation and large language models. *arXiv preprint arXiv:2403.01616*.

Anton Ehrmanntraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. New encoders

for german trained from scratch: Comparing moderngbert with converted llm2vec models.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Emotion recognition for vietnamese social media text. In *Computational Linguistics*, pages 319–333, Singapore. Springer Singapore.

Tran Thai Hoa, Tran Quang Duy, Khanh Quoc Tran, and Kiet Van Nguyen. 2025. Vifactcheck: A new benchmark dataset and methods for multi-domain news fact-checking in vietnamese. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 308–316.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nguyen Kiet, Tran Son, Nguyen Luan, Huynh Tin, Luu Son, and Nguyen Ngan. 2022. Vlsp 2021-vimrc challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(2).

Ada Kovács and Gábor Recski. 2025. Lettucedetect: A hallucination detection framework for rag applications. *arXiv preprint arXiv:2502.17125*.

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. *Preprint*, arXiv:1907.11692.

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

Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 415–426, Cham. Springer International Publishing.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *arXiv.org*.

Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. Vihealthbert: Pre-trained language models for vietnamese in health text mining. In *Proceedings of the Language Resources and Evaluation Conference*, pages 328–337, Marseille, France. European Language Resources Association.

Thinh Ngo, Khoa Dang, Son Luu, Kiet Nguyen, and Ngan Nguyen. 2024. VlogQA: Task, dataset, and baseline models for Vietnamese spoken-based machine reading comprehension. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1310–1324, St. Julian's, Malta. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Visobert: A pre-trained language model for vietnamese social media text processing. *Conference on Empirical Methods in Natural Language Processing*.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2025. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual nlu tasks. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 561–572.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. *Annual Meeting of the Association for Computational Linguistics*.

Raphael Scheible-Schmitt and Stefan Schweter. 2025. Sindbert, the sailor: Charting the seas of turkish nlp. *Preprint*, arXiv:2510.21364.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2025. Neodictabert: Pushing the frontier of bert models for hebrew. *arXiv.org*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism. *Preprint*, arXiv:1909.08053.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Issa Sugiura, Kouta Nakayama, and Yusuke Oda. 2025. llm-jp-modernbert: A modernbert model trained on a large-scale japanese corpus with long context length. *arXiv.org*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Cong Dao Tran, Nhut Huy Pham, Anh Tuan Nguyen, Truong Son Hy, and Tu Vu. 2023. ViDeBERTa: A powerful pre-trained language model for Vietnamese. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1071–1078, Dubrovnik, Croatia. Association for Computational Linguistics.

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dien Dinh. 2024. ViGLUE: A Vietnamese general language understanding benchmark and analysis of Vietnamese language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4174–4189, Mexico City, Mexico. Association for Computational Linguistics.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 named entity recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2146–2153, Online. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. Seq vs seq: An open suite of paired encoders and decoders, 2025. *URL https://arxiv.org/abs/2507.11412*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

Zeyu Zhao, Ningtao Wang, Xing Fu, and Yu Cheng. 2025. Chinese modernbert with whole-word masking. *arXiv (Cornell University)*.

A Tokenizer Evaluation

This appendix presents an intrinsic evaluation of the proposed tokenizer using established segmentation quality metrics. Following Rust et al. (2021), the study assesses tokenization efficiency using fertility and percentage of continued words (pcw),

which quantify the average number of subword tokens per word and the proportion of words split into multiple subwords, respectively. Lower values indicate more efficient segmentation.

Multiple tokenizers were trained with identical configurations but varying training data sizes (5M, 10M, 15M, and 20M documents). For comparison, widely used Vietnamese tokenizers are included: PhoBERT (Nguyen and Nguyen, 2020), ViBERT (Bui et al., 2020), ViDeBERTa (Tran et al., 2023), and ViSoBERT (Nguyen et al., 2023), which differ in vocabulary size and training methodology. Evaluation was conducted on two held-out text collections independent of the tokenizer training data. All texts were normalized and tokenized using each tokenizer's default preprocessing pipeline.

Table 6: Tokenizer evaluation on FineWiki and FineWeb-2 datasets using fertility and percentage of continued words (PCW). Lower values indicate more efficient segmentation.

Tokenizer	FineWiki		FineWeb-2		Vocab Size
	Fertility	PCW	Fertility	PCW	
5m	1.4751	0.2501	1.1458	0.1121	32K
10m	1.5058	0.2505	1.1454	0.1118	32K
15m	1.5257	0.2507	1.1458	0.1118	32K
20m	1.5823	0.2510	1.1460	0.1119	32K
PhoBERT	1.6425	0.2482	1.2082	0.1232	64K
ViDeBERTa	1.7651	0.2787	1.1947	0.1421	128K
ViBERT	1.9531	0.5304	1.5685	0.4541	38K
ViSoBERT	2.0509	0.2544	1.1925	0.1290	15K

Table 6 presents results on both evaluation sets. The proposed tokenizer variants achieve lower fertility than existing Vietnamese tokenizers while maintaining comparable or superior pcw, despite using a smaller or similar vocabulary size (32K). Across both datasets, varying the tokenizer training data size results in only marginal differences in fertility and pcw. Performance saturates early, with tokenizers trained on 5M documents achieving nearly identical segmentation efficiency to those trained on 20M documents.

These results demonstrate that the proposed tokenizer exhibits more efficient subword segmentation and effective lexical coverage compared to existing Vietnamese tokenizers. The minimal performance variation across training data sizes provides empirical support for training tokenizers on a stable subset of the corpus rather than the full pre-training data, substantially reducing tokenizer training cost while maintaining effective segmentation properties. Based on these findings, the 15M document variant was selected for model pre-training, balancing segmentation quality with training efficiency. The findings align with recent observations in SindBERT (Scheible-Schmitt and Schweter, 2025) and LLäMmlein (Pfister et al., 2025), justifying the tokenizer design choices adopted in this work.

B Pretraining details

This appendix reports the architecture specification and pretraining hyperparameters used in our experiments, providing a single reference that supports reproducibility and avoids duplicating implementation details in the main text.

Table 7 reports the complete ModernBERT-style architectural configuration used for HuITBERT$_{tiny,base}$, including depth/width and attention-related settings.

Table 7: ModernBERT model design used in this work.

	Tiny	Base
Vocabulary size	32,064	32,064
Unused tokens	64	64
Layers	10	22
Hidden size	384	768
Transformer block	PreNorm	PreNorm
Activation	GeGLU	GeGLU
Linear bias	False	False
Attention	RoPE	RoPE
Attention heads	12	12
Global attention	every three layers	every three layers
Local attention window	128	128
Intermediate size	1,152	1,152
Normalization	LayerNorm	LayerNorm
Norm ϵ	10^{-5}	10^{-5}
Norm bias	False	False
RoPE θ	160,000	160,000
Local-attn RoPE θ	10,000	10,000
Total Params	32M	134M
Embedding Params	12M	25M

Table 8 reports the sequence-length statistics of the two chunked corpora used in the stage-wise pretraining schedule. We train each stage for four epochs, so the effective token budgets in Table 9 correspond to approximately four passes over the stage-specific training splits.

Table 9 provides the stage-wise pretraining configuration for both model sizes, including the total training tokens, (micro) batch sizes, learning-rate schedules, optimizer hyperparameters, and initialization. Unless otherwise noted, Section 3 refers to this table for the exact numerical settings.

Table 8: Sequence-length statistics of the pretraining corpora.

	Min	Max	Mean	Median
Stage 1	6	1,024	712.70	756.00
Stage 2	132	4,096	1,912.36	1,574.00

C Reproducibility

To demonstrate the reproducibility and practical feasibility of our training pipeline, we report empirical throughput and efficiency statistics collected from both training stages for HuITBERT$_{base}$, trained on a single node with 8× NVIDIA RTX A6000 GPUs, and HuITBERT$_{tiny}$, trained on a single node with 4× NVIDIA RTX A6000 GPUs.

For exact reproduction, all runs use the same setup as Section 3 and Appendix B: automatic mixed precision in bfloat16, masking probability 0.30 for training (0.15 for evaluation), padding-token exclusion from the loss, sequence packing during training, and a fixed random seed of 17. Stage 1 performs the main pretraining run at shorter context, while Stage 2 continues training at 4,096 tokens with the same backbone and updated context settings. The global and micro-batch sizes in Table 9 fully determine gradient accumulation for each model size.

Table 10 summarizes the wall-clock training logs for each model and stage. We report aggregate throughput measured from real training logs under the multi-GPU setup above. These values depend on our implementation, packing behavior, and software stack, and should therefore be interpreted as reproducibility-oriented system measurements rather than universal hardware-independent limits.

The public ModernBERT[14] repository further supports reproducibility by releasing the source code, training configurations, and environment setup underlying this model family.

In practice, we recommend reproducing in three steps: (i) pretrain Tiny for Stage 1 and Stage 2 using the exact token budgets and optimizer settings in Table 9; (ii) initialize Base Stage 1 from the Stage 1 Tiny checkpoint using weight tiling from the middle, run Base Stage 1 with the Base-specific batch and learning-rate values, and then continue to Base Stage 2 from that Base Stage 1 checkpoint; and (iii) verify throughput against Table 10 using averaged training logs after warmup. Small variations in tokens/s are expected across CUDA, driver, and PyTorch versions, but the relative comparison between Stage 1 and Stage 2 should remain stable under the same implementation.

[14] https://github.com/AnswerDotAI/ModernBERT

D Training Configuration for Downstream Tasks

Table 11 summarizes the task-specific learning rates used for HuITBERT variants and other baselines, together with the evaluation metric for each main-benchmark downstream task. Table 12 reports the corresponding task-specific settings for the long-context evaluation tasks.

For the main-benchmark classification and structured prediction tasks, we report weighted F1, defined as

$$F1_{\text{weighted}} = \sum_{c=1}^{C} \frac{n_c}{N} F1_c$$

where

$$F1_c = \frac{2P_cR_c}{P_c + R_c}, \quad P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}.$$

and n_c is the number of gold instances in class c, with $N = \sum_{c=1}^{C} n_c$.

For ViFactCheck, we report Macro-F1, defined as

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} F1_c,$$

using the same class-wise definitions of precision, recall, and F1 above.

For ViQuAD 2.0 and VlogQA, we report Exact Match (EM), defined as the fraction of examples for which the normalized predicted answer exactly matches the normalized gold answer:

$$\text{EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\,T(\hat{a}_i) = T(a_i)\,],$$

where $T(\cdot)$ denotes answer normalization and tokenization.

For the main benchmark tasks, we use 10 training epochs, a batch size of 32, a maximum sequence length of 512, and dropout of 0.1. We also use a warmup ratio of 0.1, weight decay of 0.01, and gradient clipping with a max norm of 1.0; early stopping is applied with a patience of 2 epochs. For each experiment, we run downstream

Table 9: ModernBERT-style training settings used for HuITBERT pretraining.

	Stage 1: Pretraining phase		Stage 2: Context extension	
	Tiny	Base	Tiny	Base
Training tokens	260B	260B	90B	90B
Global batch size	2048	1024	1024	512
Microbatch size	128	64	48	16
Warmup tokens	2B	1.5B	–	–
Peak learning rate	8×10^-4	5×10^-4	5×10^-4	3×10^-4
Schedule	warmup–stable–decay	warmup–stable–decay	constant (no warmup)	constant (no warmup)
Weight decay	10^-5	10^-5	10^-5	10^-5
Initialization	Megatron-LM init	Tiny	continue from Stage 1	continue from Stage 1
Optimizer	StableAdamW	StableAdamW	StableAdamW	StableAdamW
β	$[0.9,\ 0.98]$	$[0.9,\ 0.98]$	$[0.9,\ 0.98]$	$[0.9,\ 0.98]$
ϵ	10^-6	10^-6	10^-6	10^-6

Table 10: Training throughput statistics across the two training stages. All values are averages measured from real-time training logs.

Metric	Base		Tiny	
	Stage 1	Stage 2	Stage 1	Stage 2
Micro-batch size	64	16	128	48
Packing efficiency	99.97%	99.42%	99.99%	99.97%
Batch service time (s)	2.04	4.21	5.27	2.04
Tokens/second	558,977	486,480	1,166,810	490,186

Table 11: Task configuration for downstream fine-tuning.

Task	Learning Rate			Metric
	HuITBERT$_{tiny}$	**HuITBERT**$_{base}$	**Other**	
PhoNER	5×10^{-5}	5×10^{-5}	2×10^{-5}	Weighted F1
UIT-VSMEC	5×10^{-5}	4×10^{-5}	1×10^{-5}	Weighted F1
ViHSD	4×10^{-5}	4×10^{-5}	1×10^{-5}	Weighted F1
ViNLI	5×10^{-5}	5×10^{-5}	1×10^{-5}	Weighted F1
ViQuAD 2.0	3×10^{-5}	3×10^{-5}	3×10^{-5}	EM
VLSP 2013 POS	5×10^{-5}	5×10^{-5}	2×10^{-5}	Weighted F1
VSFC Sentiment	3×10^{-5}	5×10^{-5}	2×10^{-5}	Weighted F1
VSFC Topic	2×10^{-5}	5×10^{-5}	2×10^{-5}	Weighted F1
VToC	5×10^{-5}	5×10^{-5}	2×10^{-5}	Weighted F1

fine-tuning three times with random seeds 10, 42, and 100 to improve the reliability of the results. Optimization is performed with AdamW using betas $(0.9, 0.999)$ and epsilon 1×10^{-8}, together with a linear learning-rate schedule after warmup. For the long-context tasks, Table 12 lists the task-specific maximum length, dropout, learning rate, and metric.

Table 12: Task configuration for long-context downstream evaluation.

Task	Setting	**HuITBERT**$_{tiny}$		**HuITBERT**$_{base}$	
		Stage 1	**Stage 2**	**Stage 1**	**Stage 2**
ViFactCheck	Max Length	1024	4096	1024	4096
	Dropout	0.3	0.3	0.3	0.3
	Learning Rate	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}
	Metric	Macro-F1	Macro-F1	Macro-F1	Macro-F1
VlogQA	Max Length	1024	4096	1024	4096
	Dropout	0.3	0.3	0.3	0.3
	Learning Rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
	Metric	EM	EM	EM	EM

E Downstream Task Details

This appendix provides brief descriptions of the downstream datasets used in the main benchmark.

E.1 Single Sentence

VToC (Tran et al., 2024) is a single-sentence classification dataset derived from Vietnamese online news articles. Each instance corresponds to an abstract sentence extracted from a news article and is labeled with its associated topic. The dataset covers 15 predefined categories, including Business, Education, Entertainment, Health, Science, Sports, Travel, and World. VToC aims to assess a model's ability to capture topical semantics and perform fine-grained topic classification on Vietnamese news text.

VSFC (Nguyen et al., 2018) is a Vietnamese text classification dataset designed for sentiment and topic classification of student feedback collected from end-of-semester surveys at a Vietnamese university. The corpus comprises over 16,000 human-annotated sentences with both sentiment polarity

labels (positive, neutral, negative) and topic labels (e.g., lecturer, curriculum, facility, other). An annotation scheme with discrete class labels was applied, and inter-annotator agreement rates of 91.20% and 71.07% were reported for the sentiment and topic tasks, respectively. Each instance consists of a feedback sentence paired with its corresponding sentiment and topic labels to support supervised evaluation of Vietnamese NLP models.

VSMEC (Ho et al., 2020) is a Vietnamese emotion recognition dataset designed for sentence-level classification of social media text. The dataset comprises 6,927 human-annotated comments collected from Facebook, each labeled with one of seven emotion categories: enjoyment, sadness, anger, fear, disgust, surprise, and other. An explicit single-label annotation scheme is applied to each sentence, with inter-annotator agreement exceeding 80%, ensuring label consistency. Each instance consists of a raw Vietnamese social media comment paired with a discrete emotion label, supporting supervised evaluation of emotion classification models in informal Vietnamese text.

ViHSD (Luu et al., 2021) is a large-scale Vietnamese hate speech detection dataset formulated as a sentence-level text classification task on social media content. The dataset contains 33,400 human-annotated comments collected from Facebook and YouTube, covering diverse domains such as entertainment, social issues, and politics. Each instance is labeled with one of three mutually exclusive categories: clean, offensive, or hate following explicit annotation guidelines and majority voting to ensure label reliability. The corpus is split into training, development, and test sets and serves as a benchmark for supervised evaluation of Vietnamese hate speech detection models.

E.2 Sentence Pair

ViNLI (Huynh et al., 2022) is an open-domain Vietnamese natural language inference dataset comprising 30,376 human-annotated premise-hypothesis sentence pairs. The data are extracted from more than 800 online news articles covering 13 diverse topics, including Business, Health, Law, Science, and World news. Each sentence pair is annotated with one of four inference labels: entailment, contradiction, neutral, and other. The dataset is designed to support robust evaluation of Vietnamese NLI models in realistic, news-based language understanding scenarios.

E.3 Machine Reading Comprehension

ViQuAD2.0 (Kiet et al., 2022) is a Vietnamese machine reading comprehension dataset designed for span-extraction question answering. The dataset comprises approximately 36,000 human-annotated question-answer-passage triples, including both answerable and unanswerable questions. It is constructed by extending UIT-ViQuAD 1.0 with more than 12,000 newly generated unanswerable questions, following a controlled transformation and validation process inspired by SQuAD 2.0. Each instance consists of a passage, a question, and either an answer span extracted from the passage or an explicit indication that the question is unanswerable. UIT-ViQuAD 2.0 covers diverse topics drawn from Vietnamese textual sources and serves as a challenging benchmark for evaluating Vietnamese machine reading comprehension and question answering systems.

E.4 Structured Prediction

PhoNER (Truong et al., 2021) is a Vietnamese named entity recognition dataset designed for sequence labeling in the COVID-19 domain. The dataset comprises 10,027 manually annotated sentences collected from reputable Vietnamese online news sources, containing 34,984 entity mentions. It adopts a BIO tagging scheme with ten domain-specific entity types covering patient identifiers, personal attributes, locations, organizations, medical conditions, transportation, and temporal expressions. Each instance consists of a sentence annotated at the entity-span level, enabling supervised evaluation of Vietnamese NER models in a specialized public health context.

VLSP-2013 POS Tagging [15] is a standard benchmark dataset for Vietnamese part-of-speech (POS) tagging. The dataset comprises 27,870 manually annotated sentences for training and development, split into approximately 27,000 for training and 870 for development, along with 2,120 test sentences, collected from Vietnamese news articles. It employs a rich tagset of 18 lexical categories (e.g., noun, verb, adjective, pronoun, classifier) and distinct tags for punctuation and special characters.

F Scientific Artifacts

The licenses for all models, datasets, and software used in this paper are listed below: ModernBERT (Apache License 2.0), Sentence-BERT (Apache

[15] https://vlsp.org.vn/vlsp2013/eval

License 2.0), FineWiki (Creative Commons Attribution Share Alike 4.0), FineWeb-2 (Open Data Commons License Attribution family), mmBERT (MIT License), SEA-LION-ModernBERT (MIT License), Vietnamese-Embedding (Apache License 2.0), mBERT (Apache License 2.0), ViBERT (Apache License 2.0), PhoBERT (MIT License), VlogQA (Creative Commons Attribution Share Alike 4.0), ViFactCheck (MIT License), ViHSD (MIT License), VToC (CC BY-NC-ND 4.0), XLM-RoBERTa (MIT License), ViSoBERT (MIT License), mMARCO (Apache License 2.0), Vietnamese Bi-Encoder (Apache License 2.0), HuggingFace Transformers (Apache License 2.0), FlashAttention (BSD 3-Clause License).