# Pretraining and Benchmarking Modern Encoders for Latvian

**Arturs Znotins**
Institute of Mathematics and Computer Science, University of Latvia
29 Raina bulv., Riga, LV-1459, Latvia
arturs.znotins@lumii.lv

## Abstract

Encoder-only transformers remain essential for practical NLP tasks. While recent advances in multilingual models have improved cross-lingual capabilities, low-resource languages such as Latvian remain underrepresented in pretraining corpora, and few monolingual Latvian encoders currently exist. We address this gap by pretraining a suite of Latvian-specific encoders based on RoBERTa, DeBERTaV3, and ModernBERT architectures, including long-context variants, and evaluating them across a diverse set of Latvian diagnostic and linguistic benchmarks. Our models are competitive with existing monolingual and multilingual encoders while benefiting from recent architectural and efficiency advances. Our best model, lv-deberta-base (111M parameters), achieves the strongest overall performance, outperforming larger multilingual baselines and prior Latvian-specific encoders. We release all pretrained models and evaluation resources to support further research and practical applications in Latvian NLP.

## 1 Introduction

Large language models have shifted much of NLP research toward decoder-only architectures. However, encoder-only Transformers remain fundamental to practical natural language processing. They provide strong contextual representations for classification, sequence labeling, and span-based prediction, where systems require reliable token- and sentence-level outputs rather than free-form generation (Weller et al., 2025). Furthermore, encoders are a key component of modern retrieval-augmented generation (RAG) pipelines, serving as the backbone for retrievers and embedding models. Encoder models are also typically more compute- and memory-efficient than autoregressive LLMs, making them well suited for high-throughput and latency-sensitive use-cases.

While multilingual encoders like mBERT and XLM-R enable robust cross-lingual transfer, scaling to a diverse set of languages introduces a capacity-–coverage trade-off (Conneau et al., 2020). Because parameters and vocabularies are shared across languages with highly imbalanced pretraining corpora, low-resource languages often suffer from diluted capacity and insufficient training signals, leading to suboptimal representations. Monolingual pretraining enables better language-specific tokenization and, when sufficient unlabeled data is available, often yields stronger downstream performance (Wang et al., 2020; Rust et al., 2021). Therefore, monolingual encoders can provide a more efficient route to high-quality representations for a target language than relying on a shared multilingual model.

An open question in encoder design concerns the trade-offs among recent architectural choices. ModernBERT (Warner et al., 2025) emphasizes throughput and long-context support with modern attention implementations, whereas DeBERTaV3 (He et al., 2023) utilizes disentangled attention and replaced-token detection (RTD). However, it remains unclear how these design differences translate to low-resource, monolingual settings.

In this work, we introduce a suite of Latvian pretrained encoder models and a unified evaluation benchmark. Using a large-scale Latvian corpus, we pretrain RoBERTa-, DeBERTaV3-, and ModernBERT-based encoders under comparable recipes, including long-context variants up to 8,192 tokens, and evaluate them across a diverse set of Latvian diagnostic and linguistic benchmarks. Our best lv-deberta-base model achieves the highest overall score, substantially outperforming multilingual baselines and previously released Latvian models.

Our contributions are as follows:

- We release a suite of Latvian pretrained encoders (RoBERTa, DeBERTaV3, ModernBERT) across multiple sizes, including long-context variants[1].

- We introduce a unified benchmark suite for evaluating Latvian pretrained encoder models[2].

## 2 Related Work

Since the release of BERT (Devlin et al., 2019), encoder models have evolved through improvements in both architecture and training objectives. RoBERTa demonstrated the importance of optimized training and data scaling by showing that BERT's performance can be significantly improved through longer training, more data, and refined objectives, without architectural changes (Liu et al., 2019). DeBERTa introduced disentangled attention mechanisms that separately model token content and relative position (He et al., 2021). DeBERTaV3 further improved sample efficiency by combining disentangled attention with ELECTRA-style replaced-token detection objectives, improving sample efficiency (He et al., 2023). More recently, ModernBERT (Warner et al., 2025) and NeoBERT (Breton et al., 2025) revisit encoder design with modern attention implementations and efficiency-oriented choices to enable high throughput and long-context processing.

Latvian is supported in multilingual encoders such as mBERT (Devlin et al., 2019), where it is limited to the Latvian Wikipedia subset (~25M tokens), and in larger-scale models such as XLM-R (Conneau et al., 2020) and mDeBERTaV3 (He et al., 2023), trained on CommonCrawl–derived corpora containing substantially more Latvian data (~1.2B tokens after cleaning). More recently, mmBERT (Marone et al., 2025) further scales multilingual pretraining to 3T+ tokens across 1,800+ languages and typically outperforms XLM-R on classification, embedding, and retrieval tasks. It is designed to better support a broad range of languages by gradually adding languages during training and reducing the dominance of high-resource languages.

Several monolingual or Baltic-focused encoders have also been released. LVBERT (Znotins and Barzdins, 2020) is trained on ~0.5B tokens, primarily from news, while LitLat BERT (Ulčar and Robnik-Šikonja, 2021) jointly pretrains Latvian and Lithuanian with additional English data (4.07B tokens total). The HPLT project released Latvian monolingual BERT models trained on HPLTv2, based on substantially larger cleaned Latvian text collections (3.46B tokens) (Samuel et al., 2023; Burchell et al., 2025). Overall, these models show significant improvements over multilingual encoders.

Evaluation for Latvian remains fragmented and focuses mainly on syntactic tagging and NER. EuroEval[3] (Nielsen, 2023; Saattrup Nielsen et al., 2025) recently added Latvian datasets for reading comprehension (MultiWikiQA-lv), sentiment classification (Latvian Twitter Sentiment), linguistic acceptability (ScaLA-lv), and NER (WikiANN-lv, FullStack-NER-lv) for encoder evaluation. However, there is no unified benchmark or leaderboard for Latvian encoder models.

## 3 Pretraining

### 3.1 Dataset

All encoder models are pretrained on the same Latvian text mixture summarized in Table 1. The corpus combines large-scale web crawls with curated Latvian resources to balance coverage, quality, and topical diversity.

We include Latvian subsets from FineWeb2 (Penedo et al., 2025) and HPLT-v2 cleaned (Burchell et al., 2025). Although these datasets significantly improve multilingual coverage, Latvian content often suffers from shallow language-specific filtering, since identification and quality models are typically tuned for higher-resource languages. To improve quality without introducing heavy preprocessing, we apply additional document filtering and combine these crawls with curated Latvian collections.

The remaining sources largely come from the Latvian National Corpus Collection (LNCC) (Saulite et al., 2022), including news, legal texts, comments, and other balanced materials. We further augment the mixture with a newer Latvian Wikipedia dump, newly crawled news media, tweets, academic texts, and scanned book texts. Together, these sources support both modern usage and domain variety.

All sources undergo boilerplate and low-quality text removal. We perform exact duplicate removal

[1] https://huggingface.co/collections/AiLab-IMCS-UL/latvian-text-encoders

[2] https://github.com/LUMII-AILab/latvian-encoders

[3] https://euroeval.com/

| Name | Word count (M) |
|---|---|
| **Web-scale and curated sources** | |
| FineWeb2 | 5371 |
| HPLT-v2 (cleaned) | 3460 |
| Crawled news | 1100 |
| Tweets | 453 |
| Books | 534 |
| Academic texts | 340 |
| Wikipedia | 53 |
| **Latvian National Corpus Collection (LNCC)** | |
| News portal comments (Rozukalne et al., 2021) | 642 |
| News | 513 |
| The Balanced Corpus of Modern Latvian (Levāne-Petrova et al., 2023) | 123 |
| Legal acts (Darģis, 2022) | 116 |
| Other | 100 |
| Total after filtering and deduplication | 6430 |

Table 1: Overview of curated Latvian text corpora. Word counts are in millions.

using metadata and text matching. Near-duplicate documents are removed using MinHash LSH with 5-grams and a similarity threshold of 0.7, following the FineWeb2 processing approach (Penedo et al., 2025).

To filter low-quality or out-of-domain documents, we apply heuristics and train a 5-gram KenLM language model on the Balanced Corpus of Modern Latvian. We then score documents using perplexity: those with high perplexity are discarded, as they typically correspond to noisy, non-fluent, or weakly Latvian content.

To support context extension, we additionally sample higher quality documents with diverse token lengths, prioritizing texts longer than the original 1,024-token context window. The final context extension dataset contains 2.5B tokens in total: 1B tokens from documents of at least 4,096 tokens, 1B tokens from documents ranging between 1,024 and 4,096 tokens, and 500M tokens from shorter documents. Shorter texts are included to mitigate potential performance degradation on shorter contexts (Gao et al., 2025).

In total, the final filtered dataset contains 6.43 billion words. The preprocessing pipeline is intentionally conservative, prioritizing precision over maximum data volume. We also highlight the need for a dedicated Latvian document quality scorer to further improve future corpus construction.

### 3.2 Tokenizer

For tokenization, we use the HPLTv2 Latvian tokenizer (Burchell et al., 2025), a GPT-2–style byte-level WordPiece tokenizer with a vocabulary size of 32,768. Initial experiments did not show meaningful improvements from using a larger vocabulary.

### 3.3 Architectures

We train three model architectures. All models are pretrained on 100B tokens for fair comparison, using an effective batch size of 4.2M tokens (8k sequences of length 512) with Distributed Data Parallel (DDP). All experiments were run on a single DGX machine with 8 NVIDIA 141GB GPUs.

**RoBERTa** We follow the RoBERTa recipe implemented in HuggingFace Transformers (Liu et al., 2019). We use sequence packing and apply 30% span-based masking (Joshi et al., 2020).

**DeBERTaV3** We pretrain lv-deberta-base following the CamemBERTaV2 recipe (Antoun et al., 2023, 2024) in a single training phase using the RTD objective with a masking probability of 20%. The generator has a hidden size of 256, and we use an effective batch size of 8k sequences per optimizer update.

**ModernBERT** We follow the ModernBERT training recipe (Warner et al., 2024; Weller et al., 2025) with three-stage training. We train three model variants: mini (59M parameters), base (136M), and large (377M). We use 30% span-based masking and apply a $2\times$ smaller learning rate to avoid instability near the end of later stages. Training consists of: (i) 70B tokens in a stable phase with batch size 4096 and sequence length 1024, (ii) 20B tokens for context extension with sequence length 8096, and (iii) 10B tokens for cooldown using a $(1 - \sqrt{\text{LR}})$ schedule. During the decay phase, masking is reduced to 15%. This setup enables fair comparison of checkpoints before and after applying the decay phase. ModernBERT also introduces several training optimizations, including token-level unpadding (>99% greedy packing efficiency), FlashAttention2, and RoPE positional embeddings, yielding approximately $2\times$ faster training.

## 4 Evaluation

We evaluate the pretrained Latvian encoder models under two complementary regimes: (i) EuroEval-style lightweight diagnostics for quick screening under minimal-data fine-tuning, and (ii) more in-depth evaluation on larger linguistically grounded benchmarks, including Universal Dependencies parsing and word sense disambiguation.

As baselines, we report results for existing Latvian-specific encoders and widely used multilingual pretrained models with Latvian support. Multilingual baselines include mdeberta-v3-base (He et al., 2023), xlm-roberta-large and xlm-roberta-base (Conneau et al., 2020), mmBERT-base and mmBERT-small (Marone et al., 2025), as well as the original bert-base-multilingual model (Devlin et al., 2019). Latvian-specific baselines include hplt-bert-base-lvs (Burchell et al., 2025), litlat-bert (Ulčar and Robnik-Šikonja, 2021), and lvbert (Znotins and Barzdins, 2020).

### 4.1 Lightweight Tasks

We adapt the lightweight diagnostic task suite from EuroEval (Nielsen, 2023). Specifically, we use the EuroEval subsampled Latvian benchmark datasets and follow their task definitions and data preparation procedures closely. EuroEval down-samples each dataset to 1,024 / 256 / 2,048 samples for train/validation/test, and we use the same splits. In addition, we include COPA, a small commonsense reasoning benchmark, as it fits the same minimal-data and rapid fine-tuning regime.

In the original EuroEval protocol, all encoder models are fine-tuned with ten random seeds using a learning rate of $2 \times 10^{-5}$ and early stopping patience 2, and evaluated on bootstrapped test sets. However, we found these hyperparameters to be suboptimal for Latvian, yielding unstable model rankings. Therefore, we conducted a hyperparameter search for all models over $lr \in \{1, 2, 3, 5, 10\} \times 10^{-5}$ with early stopping patience 5. For each learning rate, we trained five models with different random seeds and selected the best-performing learning rate based on the average validation performance across all metrics. We then report test-set results for the five models trained with the selected learning rate. To improve computational efficiency, we optimize padding and checkpointing, enabling the full diagnostic suite to be completed within a few GPU hours per model.

Results are reported in Table 3.

**LTEC (Twitter Sentiment).** A sentiment classification dataset of Latvian Twitter posts from the food and drinks domain (Sproģis and Rikters, 2020; Rikters et al., 2024). The original dataset contains 5,059 training and 754 test examples. Inter-annotator agreement is 70.48%. We evaluate performance using Matthews correlation coefficient (MCC) and macro-F1 (MF1).

**ScaLA (Linguistic Acceptability).** A Latvian grammatical acceptability dataset derived from the Latvian UD treebank (Nielsen, 2023). Ungrammatical sentences are generated via constrained token deletion and swapping. Results are reported with MCC and MF1.

**FSNER (Named Entity Recognition).** A Latvian NER dataset from a multilayer syntactic–semantic corpus (approximately 60% news, 20% fiction, 10% legal, 5% spoken, 5% mixed) (Gruzitis et al., 2018). The full dataset contains 11,425 samples. We report micro-F1, both excluding the MISC class (mF1$^\dagger$) and including all classes (mF1).

**WikiQA (Reading Comprehension).** A multilingual Wikipedia QA dataset with LLM-generated questions and extractive answers (Smart, 2025). The Latvian portion contains 5,000 examples. Performance is measured using token-level F1 and exact match (EM).

**COPA (Commonsense Reasoning).** A Latvian translation of the English COPA dataset[4], machine translated and manually post-edited (Skadina et al., 2025). We further post-edit entries not corrected

[4]https://huggingface.co/datasets/AiLab-IMCS-UL/copa-lv

in the original release. We use the original splits of 400 / 100 / 500 for train/validation/test. We evaluate using MCC and accuracy (ACC).

### 4.2 Universal Dependencies

To assess how well pretrained encoders support Latvian morphosyntactic modeling, we fine-tune each model on Universal Dependencies (UD) tasks: part-of-speech tagging (UPOS/XPOS), lemmatization, morphological feature tagging, and dependency parsing. We use the Latvian UD treebank from UD v2.16 (Nivre et al., 2020), which contains approximately 19,000 manually annotated sentences, and follow the official train/validation/test split.

All encoder models are evaluated within the same unified multi-task architecture. Token-level predictions are produced using shallow feed-forward classification heads with layer normalization and dropout. For dependency parsing, we employ a biaffine graph-based parser that scores head–dependent arcs and dependency relations, and we decode trees using a maximum spanning tree algorithm under a single-root constraint. We use a learning rate of $5 \times 10^{-4}$ for the task-specific heads and a ten-times smaller learning rate for the encoder, and report results averaged over five random seeds.

We report standard UD metrics computed with the official CoNLL UD evaluation script, including token-level tagging accuracy for UPOS, XPOS, UFeats, AllTags, and Lemmas, as well as dependency parsing quality measured by UAS, LAS, CLAS, MLAS, and BLEX. Results are shown in Table 4.

### 4.3 Word Sense Disambiguation

To evaluate semantic representation quality, we construct and release a manually annotated Latvian word sense disambiguation (WSD) dataset[5] based on annotated example sentences from the Latvian WordNet (Paikens et al., 2022). These corpus examples have been manually linked to specific word senses and subsenses.

We ignore subsenses and filter out lemmas that have only a single primary sense in the inventory. The final dataset contains 1,821 lemma entries with 5,459 unique senses (approximately three senses per lemma). Each sense is associated with multiple annotated example sentences, yielding 54,364 labeled instances, including target word offsets.

[5] https://huggingface.co/datasets/AiLab-IMCS-UL/wsd-lv

We split the dataset by lemma entry to prevent lexical overlap between training and evaluation sets: 500 entries are reserved for testing, 200 for validation, and the remaining entries for training.

We follow the GlossBERT formulation (Huang et al., 2019), casting WSD as a context–gloss matching task. For each instance, the model scores the compatibility between the context sentence containing the target word and a candidate gloss, and selects the highest-scoring gloss at inference time. To improve discrimination among closely related senses, we sample hard negatives from competing senses within the same lemma entry.

We evaluate performance using binary accuracy on context–sense pairs and top-1 sense selection accuracy over the full primary sense inventory. We use a learning rate of $2 \times 10^{-5}$ and report results averaged over five random seeds. Results are shown in Table 5.

| Model | Params | Vocab |
|---|---|---|
| mdeberta-v3-base | 278M | 250K |
| xlm-roberta-large | 560M | 250K |
| xlm-roberta-base | 278M | 250K |
| mmBERT-base | 307M | 256K |
| mmBERT-small | 140M | 256K |
| bert-base-multi | 178M | 120K |
| hplt-bert-base-lvs | 124M | 33K |
| litlat-bert | 151M | 84K |
| lvbert | 111M | 32K |
| lv-deberta-base | 111M | 33K |
| lv-mbert-large | 377M | 33K |
| lv-mbert-base | 136M | 33K |
| lv-mbert-mini | 59M | 33K |
| lv-roberta-base | 124M | 33K |

Table 2: Model size (millions of parameters) and vocabulary size (thousands).

## 5 Results

Tables 3–5 summarize performance across the lightweight diagnostic tasks, UD morphosyntax, and WSD semantics. Overall, the monolingual encoders introduced in this work are competitive with multilingual baselines and prior Latvian-specific models across all evaluation regimes. In particular, lv-deberta-base achieves the most consistently strong results across tasks, despite having substantially fewer parameters than larger multilingual encoders (111M vs. 560M for xlm-roberta-large, i.e., $\approx 5\times$ smaller; Table 2).

| Model | LTEC | | ScaLA | | FSNER | | WikiQA | | COPA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **MCC** | **MF1** | **MCC** | **MF1** | **mF1$^{\dagger}$** | **mF1** | **F1** | **EM** | **MCC** | **ACC** |
| *Existing multilingual pretrained models* | | | | | | | | | | |
| mdeberta-v3-base | 49.1±1.1 | 65.8±1.2 | 54.7±2.2 | 75.9±1.7 | 86.8±0.8 | 81.4±1.0 | 64.6±1.6 | 49.2±2.0 | 17.6±5.5 | 58.8±2.8 |
| xlm-roberta-large | 52.0±1.8 | 68.0±1.9 | 56.0±0.9 | 77.3±0.4 | 87.2±0.5 | 81.5±0.8 | 71.6±0.5 | 57.1±0.7 | 1.0±4.5 | 50.5±1.7 |
| xlm-roberta-base | 47.6±2.3 | 64.4±2.0 | 44.4±1.6 | 70.1±1.0 | 83.6±1.1 | 78.4±0.7 | 51.6±3.2 | 39.6±1.9 | 2.0±2.6 | 51.0±1.3 |
| mmBERT-base | 45.1±1.7 | 62.8±1.7 | 44.6±0.4 | 71.7±0.7 | 80.4±1.5 | 74.8±2.2 | 64.3±1.8 | 49.6±2.2 | 16.4±2.9 | 58.2±1.4 |
| mmBERT-small | 39.9±1.7 | 58.1±2.6 | 36.6±0.8 | 67.9±0.7 | 77.9±1.6 | 72.4±1.8 | 57.3±3.0 | 42.9±3.3 | 6.6±5.0 | 53.3±2.5 |
| bert-base-multi | 32.2±2.5 | 53.0±3.1 | 26.6±1.0 | 61.8±1.4 | 81.0±0.7 | 75.3±0.6 | 47.0±2.7 | 35.9±2.1 | 5.5±3.8 | 52.8±1.9 |
| *Existing Latvian-specific models* | | | | | | | | | | |
| hplt-bert-base-lvs | 52.5±1.4 | 68.0±1.0 | 56.2±2.3 | 74.9±1.5 | 88.5±1.3 | 83.1±1.2 | 66.6±1.4 | 52.7±1.1 | 9.8±6.9 | 54.9±3.5 |
| litlat-bert | 49.4±1.7 | 65.7±1.4 | 51.9±0.8 | 73.8±1.4 | 84.8±0.8 | 78.7±0.9 | 59.8±1.0 | 46.4±1.2 | 15.9±2.6 | 57.9±1.3 |
| lvbert | 46.0±2.3 | 63.2±2.8 | 52.9±1.7 | 75.2±1.3 | 85.2±0.9 | 79.1±0.6 | 24.5±3.0 | 14.8±2.3 | 17.1±4.7 | 58.6±2.4 |
| *This work* | | | | | | | | | | |
| lv-deberta-base | 54.2±1.3 | 69.3±0.8 | **64.2**±1.5 | **81.5**±1.1 | **89.0**±1.0 | 83.0±1.2 | 69.2±1.5 | 54.8±2.3 | 52.5±3.5 | 76.2±1.8 |
| lv-mbert-large | 49.2±1.8 | 66.1±1.4 | 59.8±2.6 | 79.2±1.6 | 85.9±1.0 | 80.3±1.5 | **73.6**±0.3 | **59.6**±0.7 | 42.0±8.5 | 71.0±4.3 |
| lv-mbert-base | 48.0±2.9 | 63.6±4.1 | 57.1±2.7 | 78.4±1.2 | 85.1±1.1 | 78.8±1.3 | 69.0±0.6 | 54.3±0.7 | 27.4±3.4 | 63.7±1.7 |
| lv-mbert-mini | 48.8±1.2 | 65.6±1.2 | 51.3±3.4 | 74.1±2.5 | 85.1±0.8 | 79.5±0.9 | 62.8±1.3 | 48.5±1.5 | 26.7±9.1 | 63.3±4.5 |
| lv-roberta-base | 50.7±1.8 | 66.7±1.6 | 59.0±2.3 | 78.6±1.6 | 88.7±0.6 | **83.6**±0.7 | 58.1±3.4 | 44.9±3.0 | 23.6±8.8 | 61.8±4.4 |
| *LLM models* | | | | | | | | | | |
| gpt-5 | **59.1**±0.7 | 71.5±0.5 | 46.5±1.1 | 72.8±0.8 | 83.0±0.6 | 72.4±0.7 | 65.1±0.9 | 46.1±1.2 | **96.9**±0.4 | **98.5**±0.3 |
| gpt-5$^{\diamond}$ | 57.8±0.8 | **72.1**±0.5 | 57.7±0.9 | 78.8±0.6 | 85.6±0.5 | 77.8±0.6 | 65.8±0.8 | 46.3±1.1 | **96.9**±0.4 | **98.5**±0.3 |
| gemma-3-27b-it$^{\diamond}$ | 51.3±1.1 | 68.3±0.7 | 51.0±2.5 | 43.0±2.4 | 40.0±1.7 | 69.7±0.9 | 61.3±1.5 | 36.8±1.8 | 92.2±2.8 | 96.1±1.4 |

Table 3: Results for sentiment classification (LTEC), linguistic acceptability (ScaLA), named entity recognition (FSNER), question answering (WikiQA), and commonsense reasoning (COPA). MCC denotes Matthews correlation coefficient; MF1 macro-F1; mF1 micro-F1; and EM exact match. $^{\dagger}$ indicates scores excluding the MISC class. $^{\diamond}$ marks few-shot prompting for LLMs, while all other LLM results are zero-shot. The best overall result in each column is highlighted in bold, and the best result within each model group is underlined.

## 5.1 Lightweight Tasks

Results for the EuroEval-style diagnostic suite are presented in Table 3. Among existing models, xlm-roberta-large, mdeberta-v3-base, and hplt-bert-base-lvs establish strong baselines, frequently appearing as the top-performing models among previous work.

The lv-deberta-base model outperforms all other encoders across nearly all tasks, except for WikiQA, where it is surpassed by larger models (lv-mbert-large and xlm-roberta-large).

The ModernBERT-based models (lv-mbert) demonstrate strong competitiveness against existing multilingual encoders. However, compared to the strongest baselines, they achieve slightly weaker results on sentiment classification (LTEC) and named entity recognition (FSNER). They also generally trail lv-roberta-base, with the exception of WikiQA, where lv-mbert-large attains the best encoder performance.

ScaLA, WikiQA, and COPA demonstrate clear improvements over all baselines for most models introduced in this work, with the exception of lv-mbert-mini. Interestingly, on LTEC, only lv-deberta-base surpasses all baselines, even though the pretraining dataset contains internet comments and tweets. For FSNER, results converge across the strongest systems, with micro-F1 scores (excluding MISC) approaching 89, suggesting that performance on this task is reaching saturation across top encoders.

COPA remains particularly challenging: most multilingual and earlier Latvian models achieve only modest performance (Matthews correlation coefficient $< 20$), suggesting a limited ability to discriminate between causal alternatives beyond simple heuristics. Notably, lv-deberta-base is the only encoder that attains a substantially higher score (52.5), indicating markedly stronger commonsense reasoning capability. The second-best result is achieved by lv-mbert-large.

We additionally compare encoder fine-tuning to strong commercial and open LLM baselines (Table 3). Despite being orders of magnitude larger, LLMs do not consistently outperform the best encoder across diagnostic tasks. One of the strongest commercial models (GPT-5) improves over lv-deberta-base on LTEC sentiment classification (MF1 71.5 vs. 69.3), but performs worse on ScaLA, FSNER, and WikiQA under both zero-shot and few-shot prompting. In contrast, LLMs dominate the COPA reasoning task, achieving near-perfect performance (e.g., MCC 96.9), whereas even the best encoder remains substantially lower.

| Model | UPOS | XPOS | UFeats | AllTags | Lemmas | UAS | LAS | CLAS | MLAS | BLEX |
|---|---|---|---|---|---|---|---|---|---|---|
| *Existing multilingual pretrained models* | | | | | | | | | | |
| mdeberta-v3-base | 98.7±0.0 | 93.8±0.1 | 97.0±0.2 | 93.5±0.2 | 97.7±0.0 | 94.4±0.3 | 92.0±0.3 | 90.5±0.2 | 86.0±0.2 | 87.9±0.2 |
| xlm-roberta-large | 98.8±0.1 | 94.2±0.2 | 97.2±0.1 | 94.0±0.2 | **98.0**±0.1 | 94.8±0.3 | 92.5±0.2 | 91.0±0.3 | 86.9±0.3 | 88.8±0.2 |
| xlm-roberta-base | 98.5±0.0 | 93.5±0.1 | 96.7±0.1 | 93.3±0.0 | 97.5±0.1 | 93.8±0.2 | 91.3±0.2 | 89.6±0.2 | 85.0±0.3 | 87.0±0.1 |
| mmBERT-base | 98.6±0.0 | 93.7±0.0 | 96.7±0.1 | 93.4±0.0 | 97.5±0.2 | 93.8±0.2 | 91.4±0.2 | 89.8±0.2 | 85.1±0.1 | 87.1±0.4 |
| mmBERT-small | 98.3±0.1 | 93.1±0.2 | 96.4±0.3 | 92.8±0.3 | 97.1±0.1 | 93.2±0.3 | 90.6±0.1 | 88.8±0.1 | 83.8±0.4 | 85.7±0.1 |
| bert-base-multi | 98.1±0.1 | 92.4±0.2 | 95.8±0.1 | 92.1±0.1 | 96.9±0.1 | 91.5±0.3 | 88.7±0.2 | 86.8±0.2 | 81.4±0.3 | 83.9±0.2 |
| *Existing Latvian-specific models* | | | | | | | | | | |
| hplt-bert-base-lvs | 98.9±0.0 | 94.3±0.2 | 97.4±0.1 | 94.1±0.1 | 97.6±0.1 | 94.6±0.2 | 92.3±0.2 | 90.8±0.2 | 86.8±0.3 | 88.0±0.4 |
| litlat-bert | 98.8±0.1 | 93.9±0.2 | 97.1±0.1 | 93.7±0.2 | 97.5±0.1 | 94.5±0.1 | 92.1±0.1 | 90.6±0.1 | 86.3±0.1 | 87.8±0.1 |
| lvbert | 98.6±0.1 | 93.2±0.1 | 96.9±0.1 | 93.0±0.1 | 96.9±0.0 | 93.6±0.2 | 91.1±0.2 | 89.6±0.2 | 85.1±0.3 | 86.2±0.1 |
| *This work* | | | | | | | | | | |
| lv-deberta-base | **99.0**±0.1 | **94.6**±0.2 | **97.7**±0.0 | **94.4**±0.3 | **98.0**±0.1 | **95.1**±0.1 | **92.9**±0.1 | **91.4**±0.1 | **87.7**±0.1 | **89.1**±0.2 |
| lv-mbert-large | 98.9±0.0 | 94.4±0.0 | 97.4±0.1 | 94.2±0.0 | 97.8±0.2 | 95.0±0.4 | 92.7±0.3 | 91.1±0.4 | 87.0±0.4 | 88.6±0.5 |
| lv-mbert-base | 98.9±0.1 | 94.3±0.1 | 97.4±0.0 | 94.0±0.1 | 97.7±0.1 | 94.7±0.0 | 92.4±0.1 | 90.9±0.2 | 86.7±0.2 | 88.1±0.3 |
| lv-mbert-mini | 98.7±0.1 | 93.6±0.0 | 97.2±0.1 | 93.3±0.1 | 97.2±0.0 | 94.1±0.1 | 91.7±0.1 | 90.1±0.2 | 85.7±0.2 | 86.9±0.2 |
| lv-roberta-base | 98.9±0.0 | **94.6**±0.1 | 97.5±0.1 | **94.4**±0.1 | 97.9±0.0 | 95.0±0.1 | 92.7±0.0 | 91.2±0.1 | 87.4±0.1 | 88.8±0.1 |

Table 4: Universal Dependencies results on the Latvian UD treebank (UD v2.16) under joint multi-task fine-tuning. Evaluation is performed using the standard CoNLL-U scoring script.

### 5.2 Universal Dependencies

UD results are reported in Table 4. On token-level tagging tasks (UPOS, XPOS, UFeats, Lemmas), the best-performing models largely saturate performance, reflecting the relatively large size and high quality of the Latvian UD treebank. As a result, differences between encoders are small for these local classification objectives.

In contrast, larger gaps emerge for the parsing-oriented metrics (LAS, CLAS, MLAS, BLEX), which better capture structural prediction quality. While the overall differences among top-performing encoders remain modest, and no model shows a decisive advantage across all metrics, lv-deberta-base achieves the strongest and most consistent results on average, improving over both multilingual baselines and prior Latvian-specific models. These gains suggest that monolingual pretraining contributes most strongly to syntactic structure prediction beyond token-level tagging.

### 5.3 Word Sense Disambiguation

WSD results are shown in Table 5. lv-deberta-base achieves the highest accuracy (78.9% sense selection accuracy and 83.6% binary context–sense matching accuracy), with consistent improvements over both multilingual models and existing Latvian encoders. lv-mbert models show clear gains with increased model size and outperform similarly sized baselines, performing comparably to lv-roberta-base. Overall, the dataset provides good separation between model performances while exhibiting relatively low variance across runs, making it a reliable and informative evaluation task.

| Model | Match Acc. | Sense Acc. |
|---|---|---|
| *Existing multilingual pretrained models* | | |
| mdeberta-v3-base | 79.6±0.6 | 73.5±0.7 |
| xlm-roberta-large | 80.9±1.1 | 76.9±0.5 |
| xlm-roberta-base | 75.3±0.6 | 65.4±0.6 |
| mmBERT-base | 76.9±0.8 | 69.3±1.0 |
| mmBERT-small | 74.1±1.3 | 62.1±0.5 |
| bert-base-multi | 71.3±0.9 | 56.5±0.5 |
| *Existing Latvian-specific models* | | |
| hplt-bert-base-lvs | 77.4±0.5 | 72.8±0.2 |
| litlat-bert | 76.2±0.8 | 67.3±0.6 |
| lvbert | 76.5±0.7 | 68.1±0.4 |
| *This work* | | |
| lv-deberta-base | **83.6**±0.2 | **78.9**±0.3 |
| lv-mbert-large | 81.7±1.2 | 77.6±0.4 |
| lv-mbert-base | 80.1±0.7 | 73.8±0.5 |
| lv-mbert-mini | 76.3±0.5 | 67.6±0.7 |
| lv-roberta-base | 78.9±0.6 | 71.4±0.6 |

Table 5: WSD results. Match Acc. denotes binary accuracy for context–sense pairs, and Sense Acc. denotes top-1 sense selection accuracy.

## 6 Conclusion

We introduced a suite of Latvian pretrained encoder models based on RoBERTa, DeBERTaV3, and ModernBERT architectures, including long-context variants supporting up to 8,192 tokens, and compiled a comprehensive set of benchmark tasks for systematic model evaluation.

Across all benchmark tasks, lv-deberta-base achieves the strongest overall performance despite its relatively small size (111M parameters). Among the diagnostic tasks, the largest relative improvement is observed in commonsense reasoning: on COPA, lv-deberta-base reaches a Matthews correlation coefficient of 50.9, whereas most multilingual and prior Latvian encoders yield only modest correlations (MCC $< 20$). On the newly introduced WSD benchmark, lv-deberta-base attains the best sense selection accuracy (78.9%) and the highest context–sense matching accuracy (83.6%).

ModernBERT models provide competitive performance while enabling substantially higher throughput and long-context processing (up to 8,192 tokens). Long-context variants generally match their short-context counterparts and yield the largest gains on extractive question answering. However, despite these practical advantages, ModernBERT models remain consistently below lv-deberta-base while achieving results comparable to or slightly lower than lv-roberta-base. Increasing model size from mini to large leads to improvements across most tasks.

Overall, our results confirm that Latvian-focused monolingual pretraining yields substantial gains over multilingual encoders across both lightweight diagnostic tasks and linguistically grounded evaluations. We release all pretrained models and evaluation resources to support reproducible Latvian NLP research and facilitate downstream applications.

## Limitations

All models in this work are trained only on Latvian data. This monolingual scope limits cross-lingual transfer and may reduce robustness on mixed-language inputs, code-switching, and multilingual settings where multilingual encoders can be advantageous.

Although the training corpus is compiled from multiple sources, it inherits biases and noise typical of web- and news-derived text. The distribution is likely skewed towards formal written registers, while conversational language, dialectal variation, and minority varieties of Latvian are underrepresented. Some documents may also contain duplicated text, boilerplate, low-quality passages, or factual errors, which can affect the resulting representations.

Our evaluation focuses on task-oriented NLU benchmarks, UD morphosyntactic modeling, and supervised word sense disambiguation. The results do not directly measure retrieval and embedding quality, robustness under domain shift, performance on specialized domains, or behavior on long-context tasks beyond those included in our benchmark suite.

## Acknowledgments

This work was funded by the EU Recovery and Resilience Facility project “Language Technology Initiative” (2.3.1.1.i.0/1/22/I/CFLA/002).

## References

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. Camembert 2.0: A smarter french language model aged to perfection. *Preprint*, arXiv:2411.08868.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2023. Data-efficient French language modeling with CamemBERTa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5174–5185, Toronto, Canada. Association for Computational Linguistics.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, John X Morris, and Sarath Chandar. 2025. Neobert: A next-generation bert. *arXiv preprint arXiv:2502.19587*.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Roberts Darģis. 2022. Corpus of Legal Acts of the Republic of Latvia (Likumi). CLARIN-LV digital library at IMCS, University of Latvia.

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.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399, Vienna, Austria. Association for Computational Linguistics.

Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens. 2018. Creation of a balanced state-of-the-art multilayer corpus for NLU. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Kristīne Levāne-Petrova, Roberts Darģis, Kristīne Pokratniece, and Viesturs Jūlijs Lasmanis. 2023. Balanced corpus of modern latvian (LVK2022). CLARIN-LV digital library at IMCS, University of Latvia.

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. *arXiv preprint arXiv:1907.11692*.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Peteris Paikens, Mikus Grasmanis, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, and Laine Strankale. 2022. Towards Latvian WordNet. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2808–2815, Marseille, France. European Language Resources Association.

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.

Matiss Rikters, Rinalds Vīksna, and Edison Marrese-Taylor. 2024. Annotations for exploring food tweets from multiple aspects. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1233–1238, Torino, Italy. ELRA and ICCL.

Anda Rozukalne, Vineta Kleinberga, and Normunds Grūzītis. 2021. Covid-19 news and audience aggressiveness: Analysis of news content and audience reaction during the state of emergency in latvia (2020–2021). In *ENVIRONMENT. TECHNOLOGY. RESOURCES. Proceedings of the International Scientific and Practical Conference*, volume 2, pages 141–147.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

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, Tallinn, Estonia. University of Tartu Library.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Baiba Saulite, Roberts Dargis, and Normunds Gruzitis et al. 2022. Latvian National Corpora Collection – Korpuss.lv. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Inguna Skadina, Bruno Bakanovs, and Roberts Darģis. 2025. First steps in benchmarking Latvian in large language models. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 86–95, Tallinn, Estonia. University of Tartu Library, Estonia.

Dan Saattrup Smart. 2025. Multiwikiqa: A reading comprehension benchmark in 300+ languages. *Preprint*, arXiv:2509.04111.

Uga Sproģis and Matīss Rikters. 2020. What Can We Learn From Almost a Decade of Food Tweets. In *In Proceedings of the 9th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2020)*, Kaunas, Lithuania.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Training dataset and dictionary sizes matter in bert models: the case of baltic languages. In *International conference on analysis of images, social networks and texts*, pages 162–172. Springer.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders. *Preprint*, arXiv:2507.11412.

Arturs Znotins and Guntis Barzdins. 2020. LVBERT: Transformer-Based Model for Latvian Language Understanding. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press.