NeoDictaBERT: Pushing the Frontier of BERT models for Hebrew

Shaltiel Shmidman[1,†], **Avi Shmidman**[1,2,‡], **Moshe Koppel**[1,2,†]
[1]DICTA / Jerusalem, Israel [2]Bar Ilan University / Ramat Gan, Israel
[†]{shaltieltzion,moishk}@gmail.com
[‡]avi.shmidman@biu.ac.il

Abstract

Since their initial release, BERT models have demonstrated exceptional performance on a variety of tasks, despite their relatively small size (BERT-base has ~110M parameters). Nevertheless, the architectural choices used in these models are outdated compared to newer transformer-based models such as Llama3 and Qwen3. In recent months, several architectures have been proposed to close this gap. ModernBERT and NeoBERT both show strong improvements on English benchmarks and significantly extend the supported context window. Following their successes, we introduce NeoDictaBERT and NeoDictaBERT-bilingual: BERT-style models trained using the same architecture as NeoBERT, with a dedicated focus on Hebrew texts. These models outperform existing ones on almost all Hebrew benchmarks and provide a strong foundation for downstream tasks. Notably, the NeoDictaBERT-bilingual model shows strong results on retrieval tasks, outperforming other multilingual models of similar size. In this paper, we describe the training process and report results across various benchmarks. We release the models[1] to the community as part of our goal to advance research and development in Hebrew NLP.

1 Introduction

Transformer-based encoder models first achieved popularity with the release of the original BERT models (Devlin et al., 2019). Since then, multiple models have been released based on the original BERT, incorporating various modifications to boost performance. Examples include RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), ALBERT (Lan et al., 2020), and CharacterBERT (Boukkouri et al., 2020).

In recent years, transformer-based decoder models such as GPT-3 (Brown et al., 2020) and Llama (Touvron et al., 2023) have risen in popularity, driving extensive research into architectural choices of transformer-based models. As a result, many changes to the original transformer architecture have become mainstream, aimed at improving scalability, extending context windows, accelerating convergence, and enhancing performance.

Yet, while decoder models steadily evolved, encoder models largely retained their original architecture. This gap led to the recent release of ModernBERT (Warner et al., 2025) and subsequently NeoBERT (Breton et al., 2025). The key motivation behind these releases was to modernize the BERT-style architecture using innovations from newer decoder models, scaling up pretraining data by an order of magnitude. Both demonstrated substantial performance improvements on downstream tasks, with NeoBERT outperforming ModernBERT on several key benchmarks.

In this paper, we present our new models for Hebrew NLP, NeoDictaBERT and NeoDictaBERT-bilingual, which adopt the NeoBERT architecture and training methodology, and which are trained on Hebrew-centric pretraining data for Hebrew language modeling. Both models natively support a context window of 4,096 tokens (8x more than the original BERT). The NeoDictaBERT model was trained solely on Hebrew data, while NeoDictaBERT-bilingual was trained on a 60–40 mix of English and Hebrew data, respectively. We observe performance gains across almost all benchmarks when compared with previous state-of-the-art Hebrew models; the most dramatic boost occurs in QA tasks, demonstrating significant improvements in semantic understanding. Additionally, we show that NeoDictaBERT-bilingual performs strongly when fine-tuned for retrieval, on both English and Hebrew benchmarks.

[1]The models are released under a CC BY-SA license: https://creativecommons.org/licenses/by-sa/4.0/

2 Approach

2.1 Architecture

As noted, we adopt the NeoBERT architecture, which introduces several key changes compared to the original BERT. In particular:

- **Normalization**: RMSNorm (Zhang and Sennrich, 2019) applied before the main sub-layer (Pre-RMSNorm), instead of the standard LayerNorm applied after the main sub-layer and residual (Post-LayerNorm).
- **Position Encoding**: RoPE (Su et al., 2023), which encodes relative positions, allowing better generalization to longer sequences.
- **Activation Function**: SwiGLU gated activation (Shazeer, 2020), replacing the non-gated GeLU (Hendrycks and Gimpel, 2023) activation.
- **Shared Embedding & Output Weights**: The embedding layer and output LM-head weights are no longer shared.

Two additional architectural changes are worth noting. First, the context window: while the original BERT models are trained initially with a 128-token window, and then extended to 512 tokens in phase 2 of the training, here we train with an initial window of 1,024 tokens, extended to 4,096 tokens in the second phase. Second, we adjust the depth-to-width ratio, increasing the model depth to 28 layers (compared to 12 in $BERT_{base}$) while retaining the original $BERT_{base}$ width of 768. This achieves the theoretically optimal ratio as suggested by Levine et al. (2020).

2.2 Tokenizer

Both models use the WordPiece tokenization method proposed by Song et al. (2021), with the default normalizers and preprocessors suggested by HuggingFace, along with Hebrew-specific modifications introduced in DictaBERT (Shmidman et al., 2023).

The `NeoDictaBERT` model uses the same tokenizer as DictaBERT, while for `NeoDictaBERT-bilingual` we train a new tokenizer on a mixture of Hebrew and English, maintaining a vocabulary size of 128,000.

2.3 Data

- **English** We use FineWeb-Edu (Lozhkov et al., 2024), an open-source, high-quality pre-training dataset for English.
- **Hebrew** We begin with the training corpus used in Shmidman et al. (2024a), and we expand it to include new Hebrew corpora which have become available since.

2.4 Training Details and Hyperparameters

Following the NeoBERT training recipe, we train our model only on the masked-language modeling task (and not on the next-sentence prediction task). In addition, we train on 20% of the tokens (as opposed to 15%) and mask the token 100% of the time (as opposed to splitting between masking, swapping to an incorrect token, and leaving the token unchanged).

Both models were trained with a batch size of 2048, using packed sequences to minimize padding and maximize data utilization. The initial training phase used a context window of 1,024 tokens and a maximal learning rate of 6e-3, followed by a second phase extending the context window to 4,096 tokens with a maximal learning rate of 1e-4. For both phases, we used the AdamW optimizer (Loshchilov and Hutter, 2019) with 500 warmup steps and cosine annealing. The models were trained using the NeMo Framework mixed-FP8 precision plugin, resulting in a 30–50% boost in training throughput.

The models were trained using the NeMo Framework (Harper et al.), a scalable generative AI framework built for researchers and developers working on large language models, optimized for large-scale model training on NVIDIA hardware..

`NeoDictaBERT` was pre-trained on 8×H100 80GB GPUs. The first phase trained on 232B tokens (5 epochs on our Hebrew data) for a total of 88 hours. The second phase trained on 46B tokens (1 epoch) for a total of 23 hours.

`NeoDictaBERT-bilingual` was pre-trained on 16×H200 141GB GPUs on NVIDIA DGX Cloud Lepton. The first phase trained on 612B tokens (60-40 English-Hebrew split) for a total of 106 hours. The second phase trained on 122B tokens for a total of 28 hours.

3 Hebrew Experiments and Results

We compare the performance of `NeoDictaBERT` and `NeoDictaBERT-bilingual` with previ-

Model	QA			Dependency		Sentiment	NER
	EM	F1	T1nls	UAS	LAS	Acc	F1
mBERT	69.08	74.32	77.01	–	–	84.21	79.11
AlephBertGimmel	61.77	67.97	71.35	–	–	89.51	85.26
DictaBERT	70.35	77.04	80.34	90.52	86.54	**89.79**	87.01
mmBERT-base	72.14	77.75	80.7	91.63	87.9	87.32	72.49
NeoDictaBERT	76.40	82.83	85.86	91.77	87.80	89.61	86.97
NeoDictaBERT-bilingual	**82.30**	**87.49**	**90.38**	**92.32**	**88.90**	89.62	**90.43**

Table 1: Performance on Hebrew Language benchmarks

Model Name	Pair Class.	Class.	STS	Retrieve	Cluster	Rerank	Summ.	Avg
ModernBERT-base	80.5	65.6	75.5	44.8	43.3	44.1	22.6	53.8
mmBERT-base	80.2	64.8	74.8	44.9	41.7	44.9	26.0	53.9
NeoDictaBERT-bilingual	82.2	67.4	74.4	44.4	37.1	42.8	24.5	53.2

Table 2: Performance on MTEB v2 English

ous Hebrew models of similar parameter size, drawing results from their respective publications when available; otherwise, we fill in the scores by fine-tuning the models ourselves. The models we compare to are mBERT (Devlin et al., 2019), AlephBertGimmel (Gueta et al., 2023), DictaBERT (Shmidman et al., 2023, 2024b), and mmBERT (Marone et al., 2025) (multilingual-ModernBERT).

Following previous publications, we tested our models on the following Hebrew benchmarks:

Dependency Parsing We follow the experimental setup of Shmidman et al. (2024b), evaluating both UAS (unlabeled accuracy score) and LAS (labeled accuracy score) on whole words in Hebrew. We use the Hebrew UD Treebank introduced by Tsarfaty (2013) and revised to match the tagging schema of IAHLT (Zeldes et al., 2022).[2]

Named Entity Recognition (NER) We train on the NEMO dataset presented by Bareket and Tsarfaty (2021), which contains nine categories and 6,220 sentences (7,713 entities). We report the token-level F1 score for each model.

Sentiment Analysis We evaluate our models on the sentiment analysis dataset presented by Amram et al. (2018), based on 12K social media comments.

Question Answering For this task, we use the HeQ dataset by Cohen et al. (2023), which contains 30K high-quality samples from the GeekTime newsfeed and Wikipedia.

The results for these tasks are displayed in Table 1. As can be seen, NeoDictaBERT and NeoDictaBERT-bilingual outperform all previous models, with a noticeable improvement in QA scores, indicating a deeper semantic understanding.

4 Retrieval Experiments and Results

We also evaluate the retrieval performance of NeoDictaBERT-bilingual. We follow the training setup used by Marone et al. (2025) and fine-tune the model for retrieval using English-only datasets.

We compare the performance of the trained model on the MTEB v2 English benchmarks (Muennighoff et al., 2023); results are shown in Table 2. The model demonstrates performance comparable to other leading models, underscoring its strength.

We further evaluate the trained model on a Hebrew-only retrieval benchmark by submitting it to the Hebrew Semantic Retrieval National Challenge.[3] We use the same code and baseline setup, replacing the embedding model (multilingual-e5-base (Wang et al., 2024)) with our fine-tuned model. Notably, even though our model was fine-tuned only on *English* data, it improves the baseline NDCG@20 score from 0.397 to 0.404, reflecting effective generalization from the English training corpus to Hebrew input texts.

[2]The dataset is publicly released and available at https://github.com/IAHLT/UD_Hebrew

[3]https://www.codabench.org/competitions/9950/

5 Conclusion

We are happy to release these models to the public to help advance research and development in Hebrew NLP.[4]

6 Acknowledgments

We would like to thank the NVIDIA DGX Cloud Lepton team for early access to the platform, and for providing us with the necessary compute to train this model.

References

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from Modern Hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dan Bareket and Reut Tsarfaty. 2021. Neural Modeling for Named Entities and Morphology (NEMO2). *Transactions of the Association for Computational Linguistics*, 9:909–928.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Junichi Tsujii. 2020. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Amir DN Cohen, Hilla Merhav Fine, Yoav Goldberg, and Reut Tsarfaty. 2023. Heq: a large and diverse hebrew reading comprehension benchmark.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2023. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all.

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Li Jason, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Koluguri Nithin, Huang Jocelyn, Fei Jia, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Sean Naren, and Boris Ginsburg. NeMo: a toolkit for Conversational AI and Large Language Models.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations (ICLR 2021)*. Also arXiv preprint arXiv:2006.03654.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. 2020. Limits to depth efficiencies of self-attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 22640–22651. Curran Associates, Inc.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

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

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Noam Shazeer. 2020. Glu variants improve transformer.

[4] NeoDictaBERT is available at https://huggingface.co/dicta-il/neodictabert NeoDictaBERT-bilingual is available at https://huggingface.co/dicta-il/neodictabert-bilingual

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024a. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew.

Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. 2024b. MRL parsing without tears: The case of Hebrew. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4537–4550, Bangkok, Thailand. Association for Computational Linguistics.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast wordpiece tokenization.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of stanford dependencies. In *Proc. of ACL*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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.

Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of EMNLP 2022*, pages 4331–4344, Abu Dhabi, UAE.

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