Modern Bert: An Application on Clinical Data

Naïl KHELIFA
ENSAE Paris
nail.khelifa@ensae.fr

Abstract

While Large Language Models (LLMs) continue to gain prominence, encoder-only transformer architectures such as BERT remain foundational tools for a wide range of non-generative downstream tasks. Among these, BERT-based models are still extensively employed for discriminative applications like Named Entity Recognition (NER), where they often match or even rival the performance of more recent, specialized LLMs. Nevertheless, many existing pipelines continue to rely on earlier iterations—most notably the original BERT—without integrating architectural and methodological improvements introduced in recent years.

In this work, we investigate ModernBERT, a modernized encoder-only transformer featuring an enhanced architecture aimed at improving downstream task performance and computational efficiency. To evaluate the effectiveness of these improvements, we perform a systematic comparison between BERT and ModernBERT on the task of NER, where encoder models are still commonly applied. Our evaluation is conducted on a domain-specific dataset comprising clinical trial eligibility criteria. This domain-specific dataset is chosen due to the significant challenges posed by linguistic variability and specialized terminology in the biomedical field and thus the crucial role played by NER in identifying and classifying medical entities within unstructured text. All code and experiment details are available in the following Github repository.

1 Introduction

Encoder-only transformer models, popularized by BERT (Bidirectional Encoder Representations from Transformers, [1]), have long served as foundational components in modern Natural Language Processing (NLP), particularly for non-generative downstream tasks such as classification, information retrieval, and Named Entity Recognition (NER). While recent advancements in Large Language Models (LLMs) such as GPT [2], GPT-3 [3], LLaMA-2 [4] and LLaMA-3 [5] have shifted attention toward generative models, encoder-only architectures remain crucial due to their favorable trade-offs in performance, inference efficiency, and scalability.

These models continue to serve as backbones in high-impact domains, including semantic search and Retrieval-Augmented Generation (RAG) systems [6], as well as for routing and moderation in agentic frameworks [7]. Notably, in tasks like NER, encoder models often rival the performance of specialized LLMs while remaining significantly more efficient [8].

Limitations of BERT-based models Despite their enduring utility, many encoder-based pipelines continue to rely on the original BERT architecture, which suffers from well-documented limitations: fixed sequence lengths (512 tokens), outdated training data, inefficient architectural design [9], and generally inefficient architectures, whether in terms of downstream performance or computational efficiency. Though recent efforts such as MosaicBERT [10], CrammingBERT [11], and AcademicBERT [12] have introduced partial updates, they have either prioritized training efficiency or retrieval performance, often neglecting improvements in classification accuracy, inference speed, and data scale.

Importance in the context of clinical data In the biomedical field, leveraging unstructured textual data has become a key challenge for advancing research, supporting clinical decision-making, and improving access to relevant medical information[13] [14] [15]. Among such texts, clinical trial eligibility criteria represent a particularly rich source of information, yet they remain difficult to process automatically due to their linguistic complexity and variability. These criteria, typically written in natural language, outline the conditions that a patient must meet to participate in a study, and often include specific medical entities such as diseases, treatments, biological measurements, or demographic characteristics.

Contributions In this work, we introduce ModernBERT [16], a fully modernized encoder-only transformer model designed to address the limitations of legacy architectures. To study the efficency of these improvements, we evaluate the performance of BERT-based models in comparison to ModernBert on a classical NER task related to eligibility criteria for clinical trials.

2 Background on BERT

Before diving into the refinements brought by ModernBERT, one needs to understand the BERT model in detail. BERT builds upon the Transformer architecture [17], but adopts only the encoder component of the Transformer stack. Its innovation lies in its ability to learn deeply bidirectional, context-sensitive representations of text by conditioning on both left and right context in all layers. At its core, BERT consists of a stack of Transformer encoder layers that leverage multi-head self-attention and position-wise feedforward neural networks. These layers are trained using self-supervised learning objectives that allow the model to learn language representations without the need for manually labeled data.

2.1 Input Representation

Given an input sequence of tokens $x = (x_1, x_2, \dots, x_n)$, each token is first mapped to a dense vector $e_i \in \mathbb{R}^d$, where d is the hidden size of the model. To incorporate information about token order and sentence structure, BERT combines three types of embeddings:

$$h_i^{(0)} = e_i + p_i + s_i$$

where:

- e_i: token embedding of the i-th token.
- p_i: positional embedding encoding the token's position in the sequence.
- s_i: segment embedding indicating whether the token belongs to sentence A or B, a mechanism used during pretraining for Next Sentence Prediction (NSP).

These embeddings are summed to form the input to the encoder stack.

2.2 Transformer Encoder Layers

The encoder in BERT comprises L identical layers, each consisting of two main sub-layers:

Multi-Head Self-Attention (MHSA)

This mechanism allows the model to jointly attend to information from different representation subspaces at different positions. Given input hidden states h, the attention mechanism computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

where:

$$Q = hW_Q, \quad K = hW_K, \quad V = hW_V$$

are linear projections of the input.

For multi-head attention with h heads, the outputs of individual heads are concatenated and linearly transformed:

$$\text{MHSA}(h) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$

where:

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$$

Position-wise Feedforward Network (FFN)

Each encoder layer also includes a fully connected feedforward network applied independently to each position:

$$\text{FFN}(h) = \text{GELU}(hW_1 + b_1)W_2 + b_2$$

Each sub-layer (MHSA and FFN) is followed by residual connections and layer normalization:

$$h' = \text{LayerNorm}(h + \text{MHSA}(h))$$

$$h^{(l+1)} = \text{LayerNorm}(h' + \text{FFN}(h'))$$

This design allows the model to preserve gradient flow and maintain stability during training.

3 ModernBERT: Architectural and Efficiency Improvements

ModernBERT builds upon the previously described architecture but introduces several significant modifications to enhance both the model's performance and computational efficiency. These changes are inspired by recent innovations in transformer architecture and optimization strategies, with particular attention to the limitations of traditional encoder-only transformers in handling longer contexts and more diverse datasets.

3.1 Architectural Enhancements

3.1.1 Bias-Free Parameterization

Following [18], a key architectural change in ModernBERT is the removal of bias terms from all linear layers except the final output projection. This reduces redundancy in the model's parameterization, making it more efficient. By minimizing the number of parameters in intermediate layers, ModernBERT improves training efficiency, reduces overfitting risk, and better utilizes available resources

3.1.2 Rotary Positional Embeddings (RoPE)

ModernBERT incorporates rotary positional embeddings (RoPE) [19] to improve positional encoding. Unlike BERT's absolute positional embeddings, which become inefficient for long sequences or varied contexts, RoPE encodes relative positions in a rotation-invariant manner. This allows RoPE to capture token distances more flexibly, improving performance on both short and long contexts. By rotating the positional encodings for each token pair, RoPE enhances the model's ability to generalize across varying sequence lengths. This enables ModernBERT to efficiently handle longer sequences without significant architectural changes, scaling effectively to larger datasets.

3.1.3 Pre-Normalization Block

ModernBERT incorporates a pre-normalization block [20], which places layer normalization [21] before each sub-layer in the transformer's attention and feed-forward blocks. The standard transformer architecture typically places layer normalization after the sub-layer, but this can sometimes lead to instability, especially as the depth of the model increases. By switching to pre-normalization, ModernBERT stabilizes the training process, ensuring that gradient flows remain consistent throughout the network. This adjustment is crucial for training deeper transformer models, where unstable gradients can affect convergence.

3.1.4 GeGLU Activation Function

For the activation function, ModernBERT replaces the original GeLU (Gaussian Error Linear Unit) with GeGLU (Gated GELU) [22], a variant that introduces a gating mechanism to control the flow of information through the network. This new activation function has been shown to enhance the non-linearity of transformer models, allowing them to capture more complex patterns in the data. The gating mechanism in GeGLU allows for dynamic control over neuron activation, providing more expressiveness compared to the original GeLU.

The GeGLU function operates by applying two linear projections to the input and then combining them through a gating mechanism:

$$\mathrm{GeGLU}(\mathbf{x}) = (\mathbf{x}W_1) \odot \mathrm{GeLU}(\mathbf{x}W_2)$$

where W_1 and W_2 are weight matrices. This approach improves the model's ability to model complex, nonlinear relationships in the data, which is critical for tasks such as named entity recognition (NER), where fine-grained distinctions need to be made between different medical entities or clinical conditions.

3.2 Efficiency Improvements

3.2.1 Alternating Global and Local Attention

A major inefficiency in transformer models, including BERT, is the quadratic complexity of the self-attention mechanism, $O(T^2)$, where T is the sequence length. This becomes costly for long sequences, such as in document classification or semantic search.

Following recent improvements to deal with long context models [23], ModernBERT addresses this by alternating attention patterns: every third layer uses global attention (attending to all tokens), while the intervening layers use local attention within a fixed-size sliding window (e.g., 128 tokens). This hybrid approach reduces complexity to $O(T \cdot w)$, where w is the window size for local attention, while preserving long-range dependencies with global attention.

3.2.2 Unpadding for Training and Inference

Another efficiency improvement in ModernBERT is the use of unpadding [24] during both training and inference. Traditional transformer models pad sequences to a fixed length for efficient batching, which leads to unnecessary computational overhead as padding tokens do not contribute to processing.

ModernBERT addresses this by removing padding tokens from input sequences during both training and inference, dynamically adjusting the sequence length for each batch. This approach reduces memory and computational costs, particularly when handling variable-length sequences, enabling higher throughput and lower latency for tasks involving long or irregular inputs.

4 Data Exploration: The Chia Dataset

As stated above, the motivation of our work is to compare the two models we previously described on a clinical dataset, for the Named Entity Recognition task. In this section, we describe the dataset of use.

Dataset Description. The Chia dataset [25] comprises 1,000 annotated eligibility criteria extracted from interventional Phase IV clinical trials listed on `ClinicalTrials.gov`. Spanning over 160,000 tokens, each word is annotated using a set of entity labels: `Condition`, `Procedure`, `Drug`, `Person`, `Observation`, `Mood`, and a catch-all `O` label for non-entities. Annotations follow the standard BIO tagging scheme, marking the beginning (B), inside (I), or outside (O) of entity spans, thereby preserving token-level boundaries for named entity recognition (NER) tasks.

Each sentence in the dataset is treated as a separate input instance. As illustrated in Figure 1, the majority of these sentences are relatively short (fewer than 10 tokens), although outliers may reach up to 350 words in length.

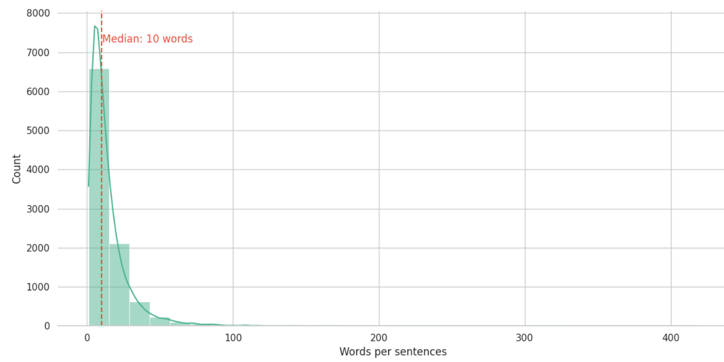


Figure 1: Token count distribution per sentence in the Chia dataset

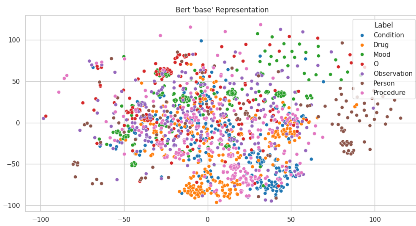
Due to the annotation strategy, a substantial proportion of tokens are labeled as `O`, as reflected in Table 1. This severe class imbalance implies that a naïve classifier predicting only the dominant label would achieve an accuracy near 75%. Hence, accuracy alone is an inadequate evaluation metric, and special consideration must be given to class-sensitive measures and training strategies that mitigate bias toward the majority class.

Label	Proportion (%)
O	74.59
Condition	13.70
Procedure	4.34
Drug	3.71
Observation	1.94
Person	1.07
Mood	0.65

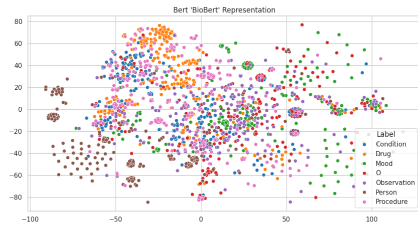
Table 1: Entity label distribution in the dataset

Excluding common stopwords, the majority of vocabulary items are medical in nature, as depicted in Figure 2. Many of these terms are rare or domain-specific, and often absent from general-domain pretraining corpora. This poses a challenge for models not exposed to biomedical language during pretraining.

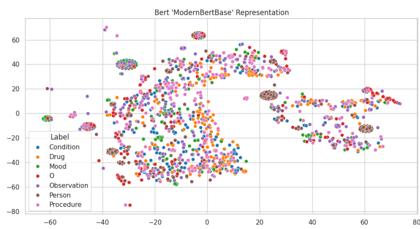
Figure 2: Word cloud of non-stopword tokens in the Chia dataset



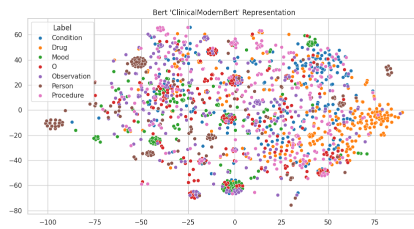
(a) BERT-Base



(b) BioBERT



(c) ModernBERT



(d) Clinical ModernBERT

Figure 3: TSNE projections of token embeddings by model and entity type

Given the clinical nature of the text, models lacking domain-specific pretraining may struggle to produce informative representations for downstream tasks. Figure 3 shows a TSNE projection of token embeddings by entity type across various models. Although none fully disentangle the entity classes, those pretrained on biomedical corpora (e.g., BioBERT, Clinical ModernBERT) exhibit more coherent and semantically relevant clustering.

This clinical NER task highlights several challenges: managing extreme label imbalance, evaluating the necessity of domain-specific pretraining, addressing dataset size limitations, and ultimately assessing whether architectures like ModernBERT can deliver both performance and efficiency gains in such specialized contexts.

5 Experimental Setup and Results

5.1 Implementation of the Model

In this section, we evaluate the performance of several transformer-based architectures on the task of Named Entity Recognition (NER) within a biomedical context. NER entails the classification of individual tokens into predefined entity categories. To ensure methodological rigor and comparability, all models are trained and evaluated under identical conditions. The dataset comprises 1,000 annotated eligibility criteria from clinical trials, partitioned into 800 samples for training, 100 for validation, and 100 for testing.

Model Selection. Model selection is based on validation performance, with early stopping applied using a patience threshold of three epochs—that is, training ceases if no improvement in validation loss is observed for three consecutive epochs. Final evaluation metrics are reported on a common test set held constant across all models.

Due to limited computational resources, exhaustive hyperparameter optimization was not feasible. Each model required between one to three hours of training time on a Tesla T4 GPU. Comprehensive training logs, implementation details, and source code are publicly available in the associated GitHub repository.

The models assessed include:

- **BERT-Base** [1]: The foundational transformer model, featuring a 768-dimensional hidden layer and 12 attention heads.

- **BioBERT-Large** [26]: Pretrained on biomedical corpora (PubMed and PMC), with an expanded architecture of 1024 hidden dimensions and 16 attention heads.
- **ModernBERT-Base** [27]: The base configuration of the ModernBERT architecture, with a 768-dimensional hidden layer and 12 attention heads.
- **Clinical ModernBERT** [28]: A variant of ModernBERT pretrained on biomedical datasets including PubMed abstracts and MIMIC-IV [29].

Loss. All models are optimized using the cross-entropy loss function, defined as:

$$\mathcal{L} = -\sum_{i=1}^{C} w_i \, y_i \log(\hat{y}_i)$$

Here, y_i denotes the true label, $\hat{y}_i$ the predicted probability for token i, and w_i the class-specific weight. When $w_i = 1$ for all classes, we refer to the loss as unweighted. Conversely, in the weighted configuration, w_i is inversely proportional to the class frequency, enhancing the influence of underrepresented classes.

Evaluation Metric. Performance is assessed using both the weighted F1 score (accounting for class distribution) and the macro F1 score (equal-weighted mean across classes).

Model	F1 Score	Macro F1 Score	Epochs
BERT	0.64	0.40	6
BioBERT	**0.87**	**0.54**	4
ModernBERT	0.81	0.43	4
Clinical ModernBERT	0.84	0.49	**3**

Table 2: Performance on the Chia dataset using unweighted cross-entropy loss

5.2 Analysis of results

As summarized in Table 2, all models achieve comparable results, with BioBERT delivering the best overall performance. Models pretrained on biomedical corpora (BioBERT and Clinical ModernBERT) consistently outperform their general-purpose counterparts, both in terms of predictive accuracy and training efficiency.

Model	F1 Score	Macro F1 Score
BERT	0.82	0.35
BioBERT	**0.87**	**0.50**
ModernBERT	0.81	0.39
Clinical ModernBERT	0.83	0.41

Table 3: Model performance after a single epoch using weighted loss

Table 3 shows that meaningful performance can be achieved even after a single epoch of training—highlighting the efficiency of pretrained transformer models in low-resource fine-tuning scenarios. BioBERT again emerges as the top-performing model, likely due to its larger architecture and richer domain-specific pretraining.

The effect of using a weighted loss function is shown in Table 4. While overall F1 scores decline slightly, this approach mitigates bias towards the dominant class and improves sensitivity to rarer entity types. The trade-off lies in reduced precision for these less frequent classes, which may or may not be acceptable depending on the application.

In conclusion, while the architectural improvements of ModernBERT do not consistently outperform BERT in this setting, models pretrained on domain-specific corpora exhibit clear advantages. Further

Model	F1 Score	Macro F1 Score	Epochs
BERT	0.64	0.38	6
BioBERT	0.64	0.41	**4**
ModernBERT	0.64	0.40	**4**
Clinical ModernBERT	**0.65**	**0.42**	**4**

Table 4: Performance using class-weighted loss function

investigations across diverse datasets and tasks are required to definitively assess the added value of architectural novelties versus pretraining strategies.

6 Conclusion and Discussion

This study presents a comparative analysis of transformer-based language models, focusing on the BERT and ModernBERT architectures, for clinical Named Entity Recognition (NER) applied to eligibility criteria of clinical trials. Despite recent innovations in model design, our empirical findings suggest that both BERT and ModernBERT yield comparable performance when fine-tuned on this domain-specific task.

More critically, the results underscore the pivotal role of pretraining data. Models pretrained on biomedical corpora—specifically BioBERT and Clinical ModernBERT—outperformed their general-purpose counterparts across all metrics. This reinforces the prevailing notion that in specialized domains such as clinical NLP, the alignment between pretraining data and downstream application domain is often more consequential than architectural sophistication.

The study highlights the continued relevance of established models like BERT when equipped with domain-adaptive pretraining. Future work should investigate scalable fine-tuning strategies, domain adaptation with fewer labeled examples, and broader benchmarking across heterogeneous medical datasets to further advance clinical language understanding.

Contribution Statement

This project was developed collaboratively by Quentin Moayedpour (quentin.moayedpour@ensae.fr) and myself (nail.khelifa@ensae.fr). The implementation was guided by official documentation and publicly available resources such as Hugging Face tutorials and community articles on transformer-based NLP models.

References

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

[2] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[3] 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. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[5] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov,

Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Con-

stable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[7] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[8] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024.

[9] Quentin G. Anthony, Jacob Hatef, Deepak Narayanan, Stella Biderman, Stas Bekman, Junqi Yin, Aamir Shafi, Hari Subramoni, and Dhabaleswar K. Panda. The case for co-designing model architectures with hardware. In *International Conference on Parallel Processing*, 2024.

[10] Sam Havens Daniel King Abhinav Venigalla Moin Nadeem Nikhil Sardana Daya Khudia Jonathan Frankle Jacob Portes, Alexander R Trott. Mosaicbert: A bidirectional encoder optimized for fast pretraining. *NeuRIPS*, 2023.

[11] Jonas Geiping and Tom Goldstein. Cramming: training a language model on a single gpu in one day. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

[12] Peter Izsak, Moshe Berchansky, and Omer Levy. How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[13] Forte. 2017 state of the clinical research industry report. 2017.

[14] J. Sedlakova, P. Daniore, A. Horn Wintsch, et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digital Health*, 2(10):e0000347, 2023. Published 2023 Oct 11.

[15] Yani Chen, Chunwu Zhang, Ruibin Bai, Tengfang Sun, Weiping Ding, and Ruili Wang. A review of medical text analysis: Theory and practice. *Information Fusion*, 119:103024, 2025.

[16] 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. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[18] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khc, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 2021.

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

[20] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[21] Jimmy Ba, Jamie Kiros, and Geoffrey Hinton. Layer normalization. 07 2016.

[22] Noam Shazeer. Glu variants improve transformer, 2020.

[23] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024.

[24] Jinle Zeng, Min Li, Zhihua Wu, Jiaqi Liu, Yuang Liu, Dianhai Yu, and Yanjun Ma. Boosting distributed training performance of the unpadded bert model, 2022.

[25] Fabricio Kury, Alex Butler, Chi Yuan, Li-Heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. Chia Annotated Datasets. 2 2020.

[26] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 2019.

[27] 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. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

[28] Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text, 2025.

[29] Alistair Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei Lehman, Leo Celi, and Roger Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 01 2023.