Transformer Models for Question Answering on ASD QA Dataset in Multilingual Setting

Alina Ryabtsev
Alina.Ryabtsev@mail.huji.ac.il

Orian Dabod
Orian.Dabod@mail.huji.ac.il

Ravid Levy
Ravid.Levy@mail.huji.ac.il

Abstract

Transformer-based question-answering (QA) models promise to enhance inclusive education by being tested and optimized with small, specific datasets before deployment. However, sociomedical research indicates these models can be unpredictable. The original study indicates that while generative QA models may misrepresent facts and produce false tokens, they can enhance system output diversity and improve user-friendliness for younger users. It also suggests that, despite the higher reliability of extractive QA models, these models might be less efficient than their generative counterparts according to the metric scores.

Our investigation further explores these findings by reproducing the existing study, evaluating QA models on the Russian dataset, and extending the research by including English-translated data. We evaluate according to the original paper (Firsanova, 2022) both extractive and generative QA models, including BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), XLM-RoBERTa (Conneau et al., 2020), ruBERT (Zmitrovich et al., 2024), and GPT-2 (Radford et al., 2019), on a custom dataset of 1,134 question-answer pairs about autism spectrum disorders (ASD) in Russian. The dataset was translated to English using the Tower-instruct (Alves et al., 2024) model, followed by a human evaluation of the translated content. An expert translated a subset of the dataset to ensure high-quality translations. The study tests the model's performance on the expert-translated data compared to the machine-translated dataset.

1 Introduction

Closed-domain question-answering (CDQA) systems can support inclusive education by providing user-friendly information about inclusion and special needs. Despite challenges due to limited datasets, Transformer-based models (Vaswani et al., 2017) have recently achieved state-of-the-art results in NLP tasks, including question answering (Chakravarti and Sil, 2021). Evaluating the safety of these models is crucial to ensure they do not misrepresent data or cause harm.

The study of Firsanova, 2022 examines the performance of Transformer-based CDQA models fine-tuned for the autism spectrum disorder domain, comparing extractive and generative QA systems using a specially compiled dataset. The dataset offers objective information about autism spectrum disorder and Asperger syndrome through reading passages paired with corresponding question-answer pairs (Firsanova, 2020) in the Russian language. In that study, decent performance was reported for testing both generative and extractive models.

In this study, we reproduced the results of the original study, and further explored the possibilities of the Transformer-based models on the same data translated in English, both automatically via a translation mechanism based on a large language model (LLM) and a subset of the data which an expert expert-translated. Two main aspects we investigated in this extension study: 1) the success of the language models in a multilingual setting, and 2) the performance of the translation mechanism vs. manual translation by an expert while performing CDQA tasks.

2 Data

The Autism Spectrum Disorder Question Answering (ASD QA) is a dataset compiled by the author of the original paper (Firsanova, 2022). It comprises 4138 question-answer records, based on 89 reading passages, from the online information resource for individuals with autism spectrum disorders (ASD) and Asperger syndrome. The dataset is inspired by the SQuAD dataset (Rajpurkar et al., 2016), although it is not synthetic yet a real-world dataset. The dataset is composed of three clusters, each containing a specific domain of information:

- General information on the autism spectrum disorder and Asperger's syndrome
- Facts about interaction and communication with people with special needs

- Practical guidelines for parents

Each record contains one question and one answer with metadata: (1) the numeric representation of an answer span in a corresponding context with its first symbol (answer start) and its last symbol (answer end), and (2) a tag is impossible that shows if a question can have a coherent answer. All the unanswerable questions in the ASD QA dataset are provided with the following plausible answers in Russian.

Our addition to the dataset is that we have translated about 50 records manually to English, and another 4138 records using automated translation by Unbabel_TowerInstruct-v0.1 (Tower) (Alves et al., 2024). This is an open-source language model that results from fine-tuning TowerBase (based on LLaMA-2 model (Touvron et al., 2023) trained on a highly multilingual dataset, with a tokenizer supporting 10 languages including Russian) on the TowerBlocks supervised fine-tuning, which is better suited for translation-related tasks. The model is trained to handle several translation-related tasks, including general machine translation (e.g., sentence and paragraph-level translation, terminology-aware translation, and context-aware translation).

Table 1 demonstrates some statistics on the various datasets, and Figure 4 demonstrates some samples from the expert-translated dataset. As it can be seen, there is no significant difference between the original dataset and the auto-translated one in terms of the length of questions, answers, or reading passages (i.e., contexts).

For both the auto-translated and the original dataset, the following split was used for training and evaluating: 70% of the dataset for training, and the rest was equally shared between the validation and test sets (15% for both). For the expert-translated dataset, since the dataset is initially small, the whole dataset was used for both training and testing.

3 Methods

The original study (Firsanova, 2022) explored the following Transformer-based architectures: M-BERT (Devlin et al., 2019), DistilBERT (D-BERT) (Sanh et al., 2020), XLM-RoBERTa (XLM) (Conneau et al., 2020), ruBERT (Zmitrovich et al., 2024), and GPT-2 (Radford et al., 2019), where all of those models are considered to be multilingual, as they were pre-trained on multilingual datasets. M-BERT, D-BERT, and XLM-R were used for the extractive QA task while the GPT-2 was used for the generative QA task. Those models were fine-tuned on the original and the translated datasets (both the Tower and the expert-translated). This was executed using PyTorch (Paszke et al., 2017) and HuggingFace interface (Wolf et al., 2020). The hyper-parameters used for training the models are described in Table 2. Overall, the optimal number of epochs decreased for the English datasets, and the batch size increased for the extractive model due to improved hardware compared to the one used in the original study.

The hardware that was used to reproduce the original results and to train the other models is the NVIDIA® GeForce RTX™ 4090 GPU with 24 GB of G6X memory. During the experiments, the fine-tuning time was recorded.

4 Results

The loss obtained on the validation set during the training iterations for each of the datasets is shown in Figure 1, Figure 2, Figure 3. The total time required for training each of the models on each dataset for the ASD QA task is presented in Table 3. Since the original hyper-parameters were used for reproducing the original study, the training times were significantly longer than training the models on the translated datasets. The expert-translated dataset is significantly shorter than the others, which achieves even lower training times.

For evaluating the success of the models on the various datasets, the following SQuAD-metrics (Rajpurkar et al., 2016) were used: *exact match (EM):* measures the percentage of predictions that match any one of the ground truth answers exactly, and the $F1$-*score:* measures the average overlap between the prediction and ground truth answer, where the predictions and the ground truth answers are bags of tokens.

The results for evaluating the trained models on the test set are presented in Table 4. The original study's results do not entirely match the ones that were reported, especially the exact match scores that were significantly lower. This could have been caused by several reasons. First, the author did not use a random seed for splitting the dataset into train, validation, and test. Second, the implementation of the metrics that were used in the original study is wrong. Lastly, according to Figure 1, it seems that the training process did not converge, which leads to mediocre performance. Nevertheless, the XLM-RoBERTa obtained the best overall performance, which corresponds to the reported in the original study. Regarding the GPT-2, the descent results that were originally reported did not reproduce in our study, while achieving lower EM scores.

For the model that was trained on the Tower translated dataset, most of the extractive models received higher F1 scores compared to the models trained on the original dataset but obtained significantly lower EM results. The XLM-RoBERTa obtained the highest results among the other models, and the GPT-2 obtained the lowest results (both EM and F1). The high F1 score might indicate that the extractive models tend to perform better (on the token level) on English data, although they are multilingual. The same statement is not necessarily true about the generative model GPT-2. Its low performance might be explained by the quality of the translation on the specific ASD dataset, which might translate sentences to be out-of-distribution compared to what GPT-2 was pre-trained on.

The last series of models was trained on the expert-translated dataset, and because of the small length of this dataset, the training data was used as a test set. However, the models performed poorly on the QA task while achieving the lowest EM and F1 scores so far. The D-BERT achieved the highest F1 score but a 0.0 EM score, and the XLM-RoBERTa achieved the highest EM score, but still significantly low compared to the previous series of models. Overall, since these models are trained on such a small dataset, they don't manage to achieve similar performance to the previous series of models.

The last experiment that was performed was to evaluate the model trained on the Tower translated dataset on the expert-translated dataset, and test the models trained on the expert-translated dataset on the Tower translated test set (see Table 5). As expected, the models trained on the expert dataset did not manage to generalize well on the Tower-translated models, achieving relatively low results both for F1 score and EM. Even though, the models translated on the Tower-translated dataset did not obtain good results or even close to those obtained on the Tower-translated test set. This might suggest though, that the manual translation is out of the distribution of the Tower-translated one, and also less consistent on the token level as the F1 scores indicate.

5 Conclusions

This study extended previous research on Transformer-based question-answering (QA) models by evaluating their performance on the Autism Spectrum Disorder (ASD) QA dataset in Russian. By creating and utilizing both expert-translated and machine-translated datasets, the study highlighted the challenges and potential of multilingual QA systems. The results indicated that while extractive models generally performed better in terms of token-level accuracy, the generative model struggled, particularly with the translated data. The findings emphasize the importance of high-quality translations and suggest that future work should focus on improving the consistency and accuracy of machine translation to enhance the performance of multilingual QA systems. Moreover, more recent generative models might perform markedly better. This research contributes valuable insights into the application of Transformer models in specialized domains, particularly in inclusive education for individuals with ASD.

6 Limitations

Naturally, having a record with content in a different language than English poses a challenge to our approach to the problem, forcing our pipeline to become language agnostic and to have multilingual capabilities. Nonetheless, in the original paper, one could detect imprecision in the amount of the data, as well as in the reported results. Finally, recall the original paper, the limited vocabulary of this literature could make models output less general and versatile answers, therefore being less expressive and accurate.

7 Ethics

Similar to any research involving medical information, the dataset and the process of this research (current and future) must respect confidentiality, be handled according to the ethics of its sources and the privacy of the data origins, and avoid any exploitation of the data for any malicious or devastate purposes.

	Original	**Tower**	**Expert**
Number of reading passages (Extractive)	89	89	15
Number of QA pairs	4138	3544	51
The maximum length of a question (word-level)	32	34	14
Minimum length of a question (word-level)	2	2	2
The maximum length of an answer (word-level)	85	94	51
Minimum length of an answer (word-level)	2	2	5
The maximum length of a reading passage (word level)	94	103	85

Table 1: Datasets statistics. Column Original = original (Russian) dataset. Column Tower = auto-translated by Unbabel_TowerInstruct-v0.1. Column Expert = expert-translated by an expert.

Parameter	**Dataset**	**M-BERT**	**D-BERT**	**XLM**	**ruBERT**	**GPT-2**
batch size	Original	1	1	1	1	2
	Tower	16	16	16	16	2 *
	Expert	16	16	16	16	2 *
learning rate	Original	3e-5	1e-5	3e-5	3e-5	3e-5
	Tower	2e-5	2e-5	2e-5	2e-5	3e-5
	Expert	2e-4	2e-4	2e-4	2e-4	3e-3
number of epochs	Original	10	20	10	10	30
	Tower	6	6	6	6	3
	Expert	6	6	6	6	6

Table 2: Hyper-parameters used for training extractive models (M-BERT, M-distilBERT, XLM-RoBERTa, ruBERT and the generative model GPT-2, for each of the datasets that the models were fine-tuned on.
* *with gradient accumulation steps*

Dataset	**M-BERT**	**D-BERT**	**XLM**	**ruBERT**	**GPT-2**
Original	19:53	27:42	27:37	20:20	3:00:03
Tower	3:44	2:34	3:53	3:44	19:26
Expert	0:17	0:12	0:26	0:17	1:15

Table 3: Time required (in minutes) for training extractive models (M-BERT, M-distilBERT, XLM-RoBERTa, ruBERT and the generative model GPT-2, for each dataset that the models were fine-tuned on.

Dataset	**metric**	**M-BERT**	**D-BERT**	**XLM**	**ruBERT**	**GPT-2**
Original	EM	21.15 %	17.41 %	**23.17 %**	19.44 %	6.84 %
	F1	47.42 %	**49.03 %**	47.15 %	46.19 %	46.17 %
Tower	EM	6.84 %	4.97 %	**11.35 %**	2.79 %	0.62 %
	F1	51.11 %	34.94 %	**68.92 %**	47.17 %	17.33 %
Expert	EM	1.96 %	0.0%	**3.92%**	1.96 %	**3.92 %**
	F1	35.40 %	**42.41 %**	26.41 %	37.9 %	18.9 %

Table 4: Results obtained on ASD QA (extractive and generative question answering), for models trained on the original dataset, Tower-translated and expert-translated one.

Trained On	**Tested on**	**metric**	**M-BERT**	**D-BERT**	**XLM**	**ruBERT**	**GPT-2**
Tower	Expert	EM	**3.92 %**	0.0 %	0.0%	1.96 %	**3.92 %**
		F1	**32.98** %	25.11 %	10.60 %	28.17 %	18.9 %
Expert	Tower	EM	2.64 %	2.48 %	4.04%	**5.44 %**	0.62 %
		F1	18.26 %	**21.24 %**	19.01 %	15.95 %	17.33 %

Table 5: Further results exploring the performance of the models trained on translated datasets and evaluated on another translated dataset.

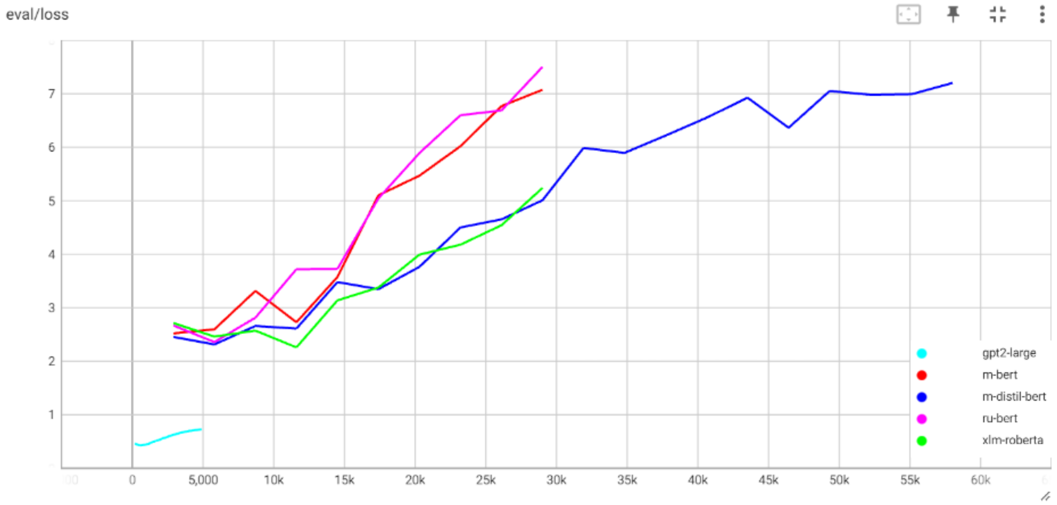


Figure 1: Chart of original loss as a function of training steps, for original dataset. The horizontal axis represents the number of the current step. The vertical axis represents the evaluation loss function of each transformer.

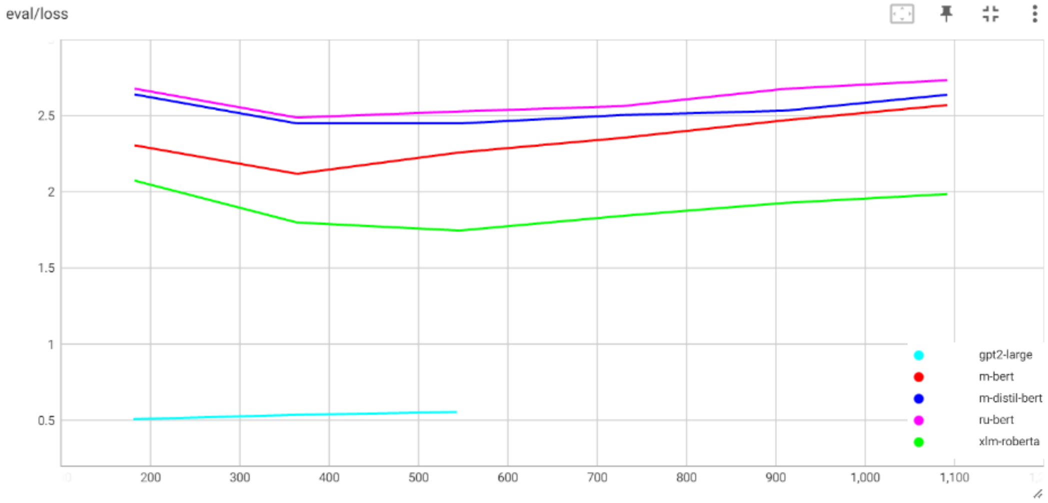


Figure 2: Chart of original loss as a function of training steps, for tower translated dataset. The horizontal axis represents the number of the current step. The vertical axis represents the evaluation loss function of each transformer.

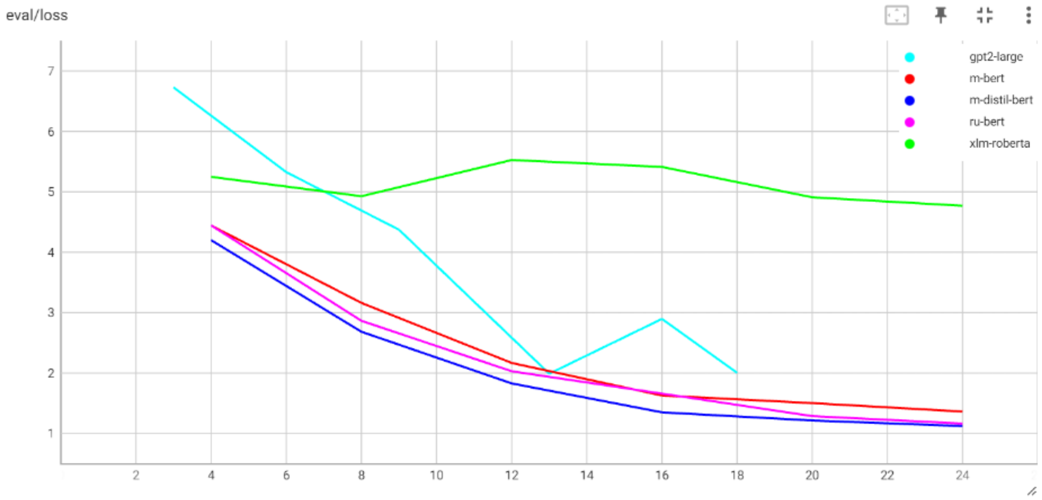
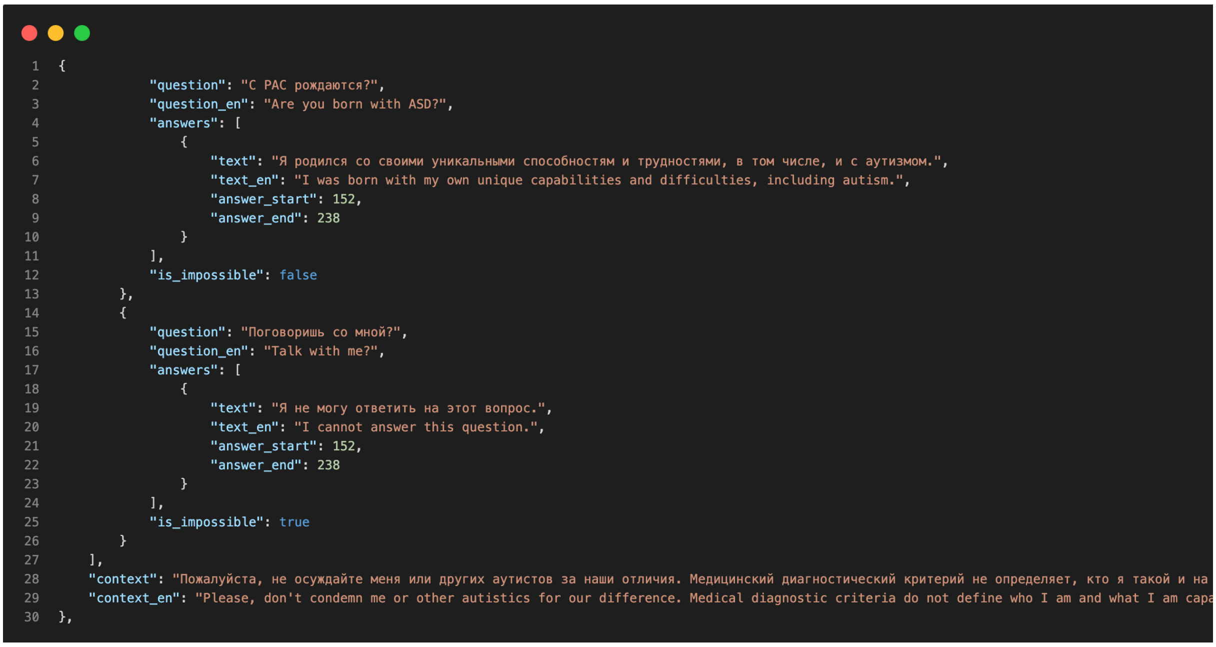


Figure 3: Chart of original loss as a function of training steps, for the expert-translated dataset. The horizontal axis represents the number of the current step. The vertical axis represents the evaluation loss function of each transformer.



```json
{
            "question": "С РАС рождаются?",
            "question_en": "Are you born with ASD?",
            "answers": [
                {
                    "text": "Я родился со своими уникальными способностям и трудностями, в том числе, и с аутизмом.",
                    "text_en": "I was born with my own unique capabilities and difficulties, including autism.",
                    "answer_start": 152,
                    "answer_end": 238
                }
            ],
            "is_impossible": false
        },
        {
            "question": "Поговоришь со мной?",
            "question_en": "Talk with me?",
            "answers": [
                {
                    "text": "Я не могу ответить на этот вопрос.",
                    "text_en": "I cannot answer this question.",
                    "answer_start": 152,
                    "answer_end": 238
                }
            ],
            "is_impossible": true
        }
    ],
    "context": "Пожалуйста, не осуждайте меня или других аутистов за наши отличия. Медицинский диагностический критерий не определяет, кто я такой и на
    "context_en": "Please, don't condemn me or other autistics for our difference. Medical diagnostic criteria do not define who I am and what I am capa
},
```

Figure 4: A snippet from the expert-translated dataset. The automatically translated dataset has also the same format. The ’en’ addition to some of the tags indicates the translation to English.

References

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Rishav Chakravarti and Avirup Sil. 2021. Towards confident machine reading comprehension.

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.

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.

Victoria Firsanova. 2020. Autism Spectrum Disorder and Asperger Syndrome Question Answering Dataset 1.0.

Victoria Firsanova. 2022. *Transformer Models for Question Answering on Autism Spectrum Disorder QA Dataset*, pages 122–133.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

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. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Vitalii Kadulin, Sergey Markov, Tatiana Shavrina, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for russian.