Exploring Large Language Models for Detecting Mental Disorders

Gleb Kuzmin[1,2,3] **Petr Strepetov**[4] **Maksim Stankevich**[3]
Natalia Chudova[3] **Artem Shelmanov**[6] **Ivan Smirnov**[3,5]
[1]AIRI [2]ISP RAS Research Center for Trusted Artificial Intelligence
[3]FRC CSC RAS [4]MIPT [5]RUDN University [6]MBZUAI
kuzmin@airi.net strepetov.pa@phystech.edu stankevich@isa.ru
nchudova@gmail.com artem.shelmanov@mbzuai.ac.ae ivs@isa.ru

Abstract

This paper compares the effectiveness of traditional machine learning methods, encoder-based models, and large language models (LLMs) on the task of detecting depression and anxiety. Five Russian-language datasets were considered, each differing in format and in the method used to define the target pathology class. We tested AutoML models based on linguistic features, several variations of encoder-based Transformers such as BERT, and state-of-the-art LLMs as pathology classification models. The results demonstrated that LLMs outperform traditional methods, particularly on noisy and small datasets where training examples vary significantly in text length and genre. However, psycholinguistic features and encoder-based models can achieve performance comparable to language models when trained on texts from individuals with clinically confirmed depression, highlighting their potential effectiveness in targeted clinical applications.[1]

1 Introduction

The problem of detecting mental disorders and patient emotions through text analysis and machine learning has been of increasing interest to researchers over the past decade (Graham et al., 2019; Zhang et al., 2022; Calixto et al., 2022; Mayer et al., 2024). In fact, advances in data science and natural language processing methods offer promising opportunities for screening, monitoring, early detection, and prevention of negative outcomes of mental disorders. Although there are studies that work with interviews (Morales and Levitan, 2016; Ringeval et al., 2017) and offline texts (Lynn et al., 2018; Stankevich et al., 2019), in most cases the material for such research comes from social media (Guntuku et al., 2017; Garg, 2023). These studies tend to focus on, but are not limited to, conditions such as depression, anxiety, stress, suicidality, post-traumatic stress disorder, and anorexia. Unsurprisingly, the methods considered for predicting mental state from text fell into traditional machine learning, using hand-crafted linguistic features, and various forms of deep learning (Zhang et al., 2022). The deep learning approach is often more accurate, especially when there are enough data samples, while traditional machine learning produces more interpretable results.

In this paper, we compare the performance of linguistic features, encoder-based models, and large language models (LLMs) on the task of identifying mental disorders. We consider two types of mental states, depression and anxiety, and several datasets in Russian that differ in text format and in the way a pathology is detected. To the best of our knowledge, this is the first comprehensive study on mental disorder detection in Russian texts, involving a wide range of techniques, including LLMs.

For our study, we involve a clinically verified dataset of essays, which contains texts written by patients with clinically diagnosed depression as well as texts from healthy volunteers (Stankevich et al., 2019). Most studies based on social media rely on self-reports, group affiliations, or questionnaire responses to determine mental health status. In contrast, only a few works use clinically validated data, which can differ substantially in quality and reliability (Chancellor and De Choudhury, 2020; Ernala et al., 2019).

This paper addresses the following research questions:

- RQ1: What is the most prominent technique for predicting depression and anxiety: traditional machine learning, encoder-based models, or recent LLMs?
- RQ2: Do models trained on a dataset of es-

[1] https://github.com/glkuzi/llm-mental-disorders-detection

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34524–34548
November 4-9, 2025

says in which depression was defined by a clinical diagnosis generalize to social media, where the depression status is defined by a questionnaire?

- RQ3: How do the LLM-generated explanations for detected depression align with the clinician's perspective?

Our main contributions are the following:

- We outperformed the existing state-of-the-art depression detection method on one dataset and established classification baselines on three previously unexamined anxiety datasets.
- We conducted a thorough comparison of various groups of models on the depression and anxiety detection tasks in Russian, which could be used by practitioners in this field for future experiments.
- We explored the transferability of models from tasks using clinical diagnoses as targets to those based on questionnaire-derived labels, aiming to mitigate the scarcity of clinically validated data in mental disorder detection.
- We evaluated and categorized LLM-generated explanations for detected depression from the point of view of clinicians, which could be used for future improvement of LLM-assisted systems in this field.

2 Related Work

2.1 Traditional and Advanced ML Methods

Researchers have employed various methods for detecting depression and anxiety across social media platforms. Tadesse et al. (2019) consider the Reddit users' dataset, comparing single and combined feature learning for depression detection. N-gram features, Linguistic Inquiry and Word Count (LIWC) dictionary features, and topics from Latent Dirichlet Allocation are considered, showcasing the effectiveness of combined features on the multilayer perceptron.

In (Shah et al., 2020), NLP methods are applied for depression detection of Reddit users based on their posts. GloVe, Word2Vec, and FastText embeddings, as well as handcrafted statistical metadata features and LIWC features, are used for text representation. The two-headed model, combining BiLSTM for embeddings and a fully connected layer for meta-features, demonstrates superior results with Word2Vec embeddings and meta-features. Additionally, the authors use Early Risk Detection Error and Latency metrics to take the time of classification into account.

Owen et al. (2020) consider depression and anxiety detection for tweets, using SVM on TF-IDF vectors and GloVe embeddings, as well as BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2020). Results show that BERT is better on a balanced dataset, while SVM excels on an unbalanced one.

Babu and Kanaga (2022) underscores the importance of emoticons in texts in sentiment analysis for depression detection. The study covers 101 publications, emphasizing the effectiveness of combining deep learning algorithms, with CNN+LSTM yielding the highest precision. Moreover, multi-class sentiment classification provides more precise results than binary and ternary classifications.

The FCL (Fasttext+CNN+LSTM) model proposed in (Tejaswini et al., 2024) outperforms LSTM and CNN models, based on GloVe and Word2vec embeddings, on datasets for depression detection with Reddit and Twitter posts.

Assessing the anxious Twitter posts caused by the COVID-19 pandemic, (Jeong et al., 2023) uses BERT trained on the Korean language, achieving strong accuracy and establishing a correlation between the anxiety index and COVID-19 waves.

The study (Ansari et al., 2023) focuses on identifying depression in social media datasets (CLPsych, Reddit, eRisk) through various text classification methods, combining sentiment lexicons with deep learning pipelines. Authors utilize sentiment lexicons with logistic regression and LSTM with attention on GloVe embeddings, comparing them and combining them into ensembles.

2.2 Large Language Models

According to the systematic review (Omar and Levkovich, 2025), most studies about depression detection focus on BERT-based models, indicating the field's early stages in adopting newer technologies like GPT-4 and Google's Gemini. However, LLMs are demonstrating significant potential in advancing depression detection systems.

The Chat-Diagnose approach (Qin et al., 2023) integrates diagnostic criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) into prompts and uses the Chain of Thoughts technique to deliver explainable diagnoses via an LLMs-augmented system based on ChatGPT/GPT-

3. This method demonstrates state-of-the-art results on Twitter and Weibo depression datasets by employing zero-shot and few-shot learning.

Another study (Hadzic et al., 2024) compares the effectiveness of fine-tuned BERT with GPT-3.5 and GPT-4 in the depression detection task. The authors use Patient Health Questionnaire-8 scores for classifying transcribed audio data from the Distress Analysis Interview Corpus, KID, and a simulated dataset. With scores separated into depressive and non-depressive groups, the zero-shot method for GPT-4 outperforms GPT-3.5 and BERT across all datasets, highlighting the potential of LLMs in depression detection.

Additionally, Wang et al. (2024) investigates depression symptom detection and severity classification using LLMs on the eRisk 2021 and eRisk 2023 datasets. Utilizing Beck's Depression Inventory to form queries related to depression symptoms and the Universal Sentence Encoder for text embeddings, the study creates two datasets containing top-1 and top-5 ranked texts for each query. LLMs fine-tuned with QLoRA are used for classification into four levels of depression severity.

The DORIS (Lan et al., 2024) system addresses the challenges of detecting depression through social media posts from the Sina Weibo Depression Dataset. The authors use GPT3.5-Turbo-1103 for annotating high-risk texts according to the DSM-5 depression scale; also, LLM is used to summarize critical information from users' historical mood records (mood courses). The final model based on XGBoost is learned on features from annotations and gte-small-zh model vector representations of post histories and mood courses and shows an improvement over the baseline.

We are the first to examine and compare three generations of the discussed models for depression and anxiety detection tasks in Russian, namely, traditional ML models, encoder-decoder models, and LLMs. Unlike other works, we used various models from each group and carefully compared the results of the models between the groups on five datasets, aiming for a general recommendation on the best models to use in practice.

3 Data

This paper considers five Russian-language datasets: 2 for depression and 3 for anxiety. Classes in all datasets were represented in the binary format: a healthy class (no signs of mental disorders) and a pathology class (depression or anxiety). The general description of the datasets used in our study is shown in Table 1.

3.1 Depression-Essays (DE)

To compile this dataset, subjects were asked to write a short essay (from 1,500 to 5,000 characters with spaces and punctuation) on the topic of "Myself, others, world" (Stankevich et al., 2019). In total, 557 essays were collected, including 110 authored by patients with clinical depression. The essays written by people with clinically validated depression were provided by the Mental Health Research Center, Moscow, Russia. The collection of essays was done on a voluntary basis, under conditions of anonymity and for research purposes only. The best-reported performance on this data reaches 73% F1 for the depression class in the cross-validation evaluation with the random forest model trained on n-grams and psycholinguistic features (Stankevich et al., 2019). For this dataset, we provided several anonymized examples in Table 21 in Appendix G.

3.2 Depression-Social Media (DSM)

The Depression-Social Media dataset was developed to support research on detecting depression in social network users (Ignatiev et al., 2022). It consists of text messages from the VKontakte platform, accompanied by results from the Russian-language adaptation of the 21-item Beck Depression Inventory (BDI) questionnaire (Beck et al., 1961). The healthy class included users with a scale score of 10 or less, and the pathological class between 30 and 63. The best-reported classification performance using textual data reaches only 65% F1 for the depression class with a logistic regression model trained on psycholinguistic features (Ignatiev et al., 2022).

We have made some changes to the original dataset. For each subject, all messages were combined for a period of 170 days prior to the date of the questionnaire screening, and the total text was limited to 6,000 characters (symbols), resulting in a median length per sample of 122 words. Such restrictions were imposed to bring the final texts from this dataset closer to the format of the essay dataset in terms of total text length for each subject and to account for the fact that the relevance of depression screening becomes less over a period of more than six months.

Name	Depression-Essays (DE)	Depression-Social Media (DSM)	Anxiety-Letter (AL)	Anxiety-Description (AD)	Anxiety-COVID Comments (AC)
Reference	(Stankevich et al., 2019)	(Ignatiev et al., 2022)	(Litvinova and Ryzhkova, 2018)		(Medvedeva et al., 2021)
Condition	Depression	Depression	Anxiety	Anxiety	Anxiety
Text format	Essay	Social media messages	Letter	Description of the picture	Short comments
Class criteria	Clinical diagnosis	BDI	HADS	HADS	SCL-90-R
# healthy samples	447	135	109	101	222
# pathology samples	110	89	93	89	191
Median words per sample	309	122	144	84	7

Table 1: Summary of datasets used in this study

3.3 Anxiety-Letter and Anxiety Picture Description (AL and AD)

For anxiety detection experiments, we use the RusNeuroPsych corpus (Litvinova and Ryzhkova, 2018). This corpus was prepared to study the relationships between a person's text, personal traits, mental status, and demographic characteristics. To compile the corpus, participants were asked to write an informal letter to a friend, provide a textual description of a picture, and complete a series of psychological questionnaires. Among these, the Hospital Anxiety and Depression Scale (HADS) (Bjelland et al., 2002) was employed to assess anxiety levels. Based on HADS scores, subjects with a score of 7 or below were assigned to the healthy class, while those with a score of 8 or higher were assigned to the pathology class. Because the corpus contains two distinct text types, it was further split into two separate datasets for the experiments. We denote them AL (Anxiety-Letter) for letters and AD (Anxiety-Description) for picture descriptions. To the best of our knowledge, no classification experiments have been performed on this data before.

3.4 Anxiety-COVID Comments (AC)

In addition, we used a corpus of subjects' comments on the COVID-19 pandemic situation (Medvedeva et al., 2021). Subjects were asked to complete a series of questionnaires and write a free-form commentary describing their attitudes towards the world situation around the pandemic and self-isolation. Among the questionnaires used was the SCL-90-R symptom questionnaire (Derogatis, 1983), the anxiety scale from which was used to form 2 groups: the healthy group, with an anxiety scale score below the 33rd percentile, and the pathology group, with an anxiety scale score above the 66th percentile. To the best of our knowledge, no classification experiments have been performed on this data before.

4 Methods

4.1 Linguistic Features

4.1.1 Psycholinguistic Features

The linguistic features used in our research were extracted using the tool described in (Smirnov et al., 2021). This tool extracts morphological, syntactic, and vocabulary parameters of texts, including various psycholinguistic coefficients. A total of 113 features were used. A detailed description of the features utilized and those extracted with this tool on our data is available in the Hugging Face repository.[2]

4.1.2 Classification Setup

As a classification baseline, we use the AutoML system auto-sklearn (Feurer et al., 2015) trained on psycholinguistic features and n-grams. The auto-sklearn classifier employs a Bayesian optimizer that considers 15 classification algorithms, 14 feature preprocessing methods, and 4 data preprocessing techniques. Additionally, the classifier utilizes a meta-learning approach and automated ensemble construction to speed up optimization.

For psycholinguistic features, we consider several feature selection methods: filter method, wrapper method (forward selection and backward elimination), and embedding method. Selected features are examined for pairwise linear correlation, and those with absolute Pearson coefficient values of more than 0.95 are deleted.

We also train models on TF-IDF vectors on unigrams and unigrams with bigrams. In addition to full vectors, we consider different subsets of features for vectors on unigrams: from 20% to 100% of the features are selected with a 20% increment. For each subset, correlated features are deleted in the same way as with psycholinguistic features.

On each set of features, we launch the classifier 6 times with different seeds and then calculate the mean and standard deviation values of the metrics

[2] https://huggingface.co/datasets/anonymizedauthor/paper_data

based on the results of the launches. The chosen AutoML models are presented in Appendix A.

4.2 Encoder-Based Models

Encoder-based classification models eliminate the need for tedious feature engineering in favor of a deep multi-layer neural architecture. As we target classification on Russian corpora, we consider models pretrained on multilingual or Russian datasets. As a simple baseline, we used a base version of multilingual BERT, a model with 110 million parameters, first introduced in (Devlin et al., 2019). Another considered baseline is RuBERT (Kuratov and Arkhipov, 2019), a pretrained Russian version of BERT-base with the same amount of parameters. We also finetuned several more recent models, such as RuBioRoBERTa (Yalunin et al., 2022), which is a RoBERTa pretrained on Russian language biomedical texts, and RuRoBERTa-large – a bigger version of RoBERTa, also pretrained on Russian language datasets. For all of the four models, we optimized hyperparameters using Bayesian optimization from the HuggingFace framework (Wolf et al., 2020).

The used hyperparameter grid and optimal parameters, along with the used checkpoint of models, are presented in Appendix A.

4.3 Large Language Models

We conducted experiments with LLMs in various settings. First of all, we evaluated models by 0-shot and 5-shot prompting, considering only normalized probabilities for tokens "0" or "1" in the first generated token, as it was done in MMLU (Hendrycks et al., 2021). We will refer to these settings as "0-shot MMLU" and "5-shot MMLU" correspondingly. Secondly, we employed 0-shot and 5-shot prompting with bigger generation lengths and matched the generated answer to one of the possible classes with string matching. Finally, we conducted the fine-tuning of the models using LoRA (Hu et al., 2022).

We selected a set of relatively small (less than 9B parameters) self-hosted open-source models, either fine-tuned for the Russian language or showing good multilingual capabilities. To the first group belongs SaigaLlama3 8B, a version of Llama 3 8B Instruct (Grattafiori et al., 2024) fine-tuned on several Russian datasets, as well as models from the Vikhr family (Nikolich et al., 2024). We used Vikhr 7B Instruct 0.4, Vikhr 7B Instruct 5.4, and Vikhr Gemma 2B Instruct. The former two are based on Mistral 7B (Jiang et al., 2023) with vocabulary adaptation for the Russian language, followed by additional pretraining and instruction tuning. The latter one is based on Gemma2 2B Instruct (Team et al., 2024), additionally trained on Russian data. We also used multilingual models, such as Gemma2 2B Instruct and Gemma2 9B Instruct, as well as Qwen2 7B Instruct (Yang et al., 2024). As we are conducting experiments on sensitive data and with private datasets, we did not consider remotely-hosted models (such as GPT-4, Claude) due to the possibility of data leakage.

The full information about used prompts, training hyperparameters, and versions of used models is presented in Appendix A.

5 Results

5.1 Classification Results on Depression and Anxiety Datasets

The data were divided into training and test samples in an 80% by 20% ratio, with stratification by the target variable reduced to binary form. All classification reports in this study show results on the test data. The classification report for the best models from each group is presented in Table 2.

5.1.1 DE Dataset

The best scores of the F1 for the pathology class and F1-macro in this experiment are achieved on the essay dataset (DE) by the fine-tuned LLM model with 88.4% F1-macro and 81.1% F1 score for the pathology class. The model trained on the linguistic features shows comparable results with 85.8% F1-macro and 77.0% F1-pathology.

In general, the three best results on the DE dataset were achieved by fine-tuned models, which can be linked to the bigger dataset size and to the longer texts in the dataset, which, in turn, are crucial for supervised fine-tuning. For all other datasets, various prompting methods without fine-tuning significantly outperform SFT and LoRA.

5.1.2 DSM Dataset

The best result achieved by the Vikhr 7B IT 0.4 model, evaluated in the 5-shot regime, with 63.69% F1-macro. The traditional machine learning methods, as well as encoder models, performed poorly with nearly 53% F1-macro on linguistic features.

Comparing the results between the DE and DSM datasets, a significant difference in classification quality can be observed. This may be due to several factors. First of all, the sample size is significantly

Corpus	Mode	Model	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
DE	SFT	Linguistic features	94.00± 0.80	95.20± 1.00	94.60± 0.80	79.30± 4.00	75.00± 3.50	77.00± 3.30	85.80± 2.10
	SFT	TF-IDF	91.60± 2.10	94.40± 2.20	93.00± 1.20	74.80± 7.00	64.40± 10.30	68.50± 6.60	80.70± 3.80
	5-shot	Vikhr 7B IT 5.4	87.06± 0.00	82.22± 0.00	84.57± 0.00	40.74± 0.00	50.00± 0.00	44.90± 0.00	64.73± 0.00
	5-shot MMLU	Gemma2 9B IT	93.11± 0.08	75.11± 0.89	83.15± 0.58	43.16± 0.85	77.27± 0.00	55.38± 0.71	69.26± 0.64
	LoRA	VikhrGemma 2B IT	94.74± 0.88	96.48± 1.62	95.59± 0.66	84.96± 5.72	78.03± 4.08	**81.13± 2.42**	**88.36± 1.52**
	SFT	RuBERT	94.45± 1.16	91.11± 1.70	92.74± 1.04	68.41± 4.04	78.03± 4.85	72.80± 3.43	82.77± 2.21
	0-shot	Gemma2 9B IT	93.33± 0.00	62.22± 0.00	74.67± 0.00	34.62± 0.00	81.82± 0.00	48.65± 0.00	61.66± 0.00
DSM	SFT	Linguistic features	62.80± 1.80	59.30± 7.40	60.70± 4.00	43.70± 2.70	47.20± 7.70	45.10± 3.80	52.90± 2.00
	SFT	TF-IDF	62.20± 2.60	51.20± 13.90	55.30± 8.50	42.90± 3.60	53.70± 11.40	46.90± 4.70	51.10± 3.50
	5-shot	SaigaLlama3 8B	100.00± 0.00	3.70± 0.00	7.14± 0.00	40.91± 0.00	100.00± 0.00	**58.06± 0.00**	32.60± 0.00
	5-shot	Vikhr 7B IT 0.4	69.19± 0.89	81.48± 0.00	74.83± 0.51	62.09± 1.10	45.56± 2.22	52.54± 1.85	**63.69± 1.18**
	5-shot MMLU	SaigaLlama3 8B	100.00± 0.00	3.70± 0.00	7.14± 0.00	40.91± 0.00	100.00± 0.00	**58.06± 0.00**	32.60± 0.00
	5-shot MMLU	Vikhr 7B IT 5.4	68.75± 0.00	81.48± 0.00	74.58± 0.00	61.54± 0.00	44.44± 0.00	51.61± 0.00	63.09± 0.00
	LoRA	Qwen2 7B IT	68.87± 6.39	72.22± 8.21	70.40± 6.75	55.33± 11.08	50.93± 10.84	52.81± 10.26	61.61± 8.32
	SFT	RuBioRoBERTa	62.10± 2.98	62.96± 6.05	62.46± 4.30	43.66± 4.43	42.59± 4.14	43.01± 3.70	52.74± 3.67
AL	SFT	Linguistic features	47.40± 21.40	36.40± 17.80	40.90± 19.20	48.50± 2.70	68.40± 14.60	56.10± 3.70	48.50± 8.20
	SFT	TF-IDF	61.50± 5.00	46.20± 3.10	52.60± 2.30	51.20± 2.50	65.80± 7.90	57.50± 4.50	55.00± 2.90
	5-shot	Gemma2 9B IT	75.00± 0.00	27.27± 0.00	40.00± 0.00	51.52± 0.00	89.47± 0.00	**65.38± 0.00**	52.69± 0.00
	LoRA	Gemma2 2B IT	60.18± 4.51	60.61± 13.55	59.66± 8.17	54.87± 7.30	53.51± 10.71	53.45± 6.06	56.56± 5.40
	SFT	RuRoBERTa	56.82± 5.65	75.00± 7.76	64.49± 5.59	52.54± 10.76	33.33± 12.77	40.17± 12.48	52.33± 8.31
	0-shot	Gemma2 9B IT	66.67± 0.00	63.64± 0.00	65.12± 0.00	60.00± 0.00	63.16± 0.00	61.54± 0.00	**63.33± 0.00**
	0-shot MMLU	Gemma2 2B IT	0.00± 0.00	0.00± 0.00	0.00± 0.00	46.34± 0.00	100.00± 0.00	63.33± 0.00	31.67± 0.00
	0-shot MMLU	Qwen2 7B IT	61.03± 0.52	85.45± 1.82	71.20± 0.99	68.73± 2.55	36.84± 0.00	47.95± 0.64	59.58± 0.82
AD	SFT	Linguistic features	50.80± 3.30	43.30± 12.50	45.30± 8.10	44.70± 2.60	51.90± 15.90	47.30± 7.40	46.30± 1.80
	SFT	TF-IDF	54.90± 15.10	37.50± 4.80	43.20± 2.50	44.60± 7.80	59.30± 21.00	50.40± 12.70	46.80± 7.10
	LoRA	Vikhr 7B IT 0.4	59.81± 4.42	51.67± 12.13	54.66± 7.59	53.47± 4.23	61.11± 11.56	56.43± 6.19	55.55± 4.55
	SFT	RuRoBERTa	57.63± 2.52	56.67± 4.71	57.07± 3.19	52.80± 2.82	53.70± 4.14	53.17± 2.84	55.12± 2.58
	0-shot	Vikhr 7B IT 5.4	83.33± 0.00	25.00± 0.00	38.46± 0.00	53.12± 0.00	94.44± 0.00	**68.00± 0.00**	53.23± 0.00
	0-shot	Qwen2 7B IT	67.25± 1.16	80.00± 0.00	73.07± 0.68	71.81± 0.76	56.67± 2.22	63.33± 1.67	68.20± 1.17
	0-shot MMLU	Vikhr 7B IT 5.4	83.33± 0.00	25.00± 0.00	38.46± 0.00	53.12± 0.00	94.44± 0.00	**68.00± 0.00**	53.23± 0.00
	0-shot MMLU	Qwen2 7B IT	70.74± 1.47	70.00± 0.00	70.36± 0.72	67.02± 0.70	67.78± 2.22	67.39± 1.44	**68.87± 1.08**
AC	SFT	Linguistic features	64.80± 16.00	45.60± 21.00	47.00± 19.60	48.70± 3.60	60.50± 20.10	52.60± 7.30	49.80± 7.80
	SFT	TF-IDF	56.30± 2.10	63.30± 4.20	59.50± 2.50	48.90± 3.30	41.70± 5.60	44.90± 4.30	52.20± 2.70
	5-shot	Qwen2 7B IT	84.89± 3.56	26.22± 5.33	39.85± 6.96	52.08± 1.72	94.74± 0.00	**67.20± 1.46**	53.52± 4.21
	5-shot MMLU	SaigaLlama3 8B	100.00± 0.00	13.33± 0.00	23.53± 0.00	49.35± 0.00	100.00± 0.00	66.09± 0.00	44.81± 0.00
	LoRA	Vikhr 7B IT 5.4	61.87± 2.07	62.96± 8.18	62.13± 4.30	55.50± 3.03	53.95± 7.25	54.33± 3.58	**58.23± 2.38**
	SFT	BERT	57.01± 5.15	62.96± 16.81	58.44± 6.53	46.07± 8.77	41.67± 22.14	41.28± 17.15	49.86± 6.84
	0-shot	SaigaLlama3 8B	66.05± 0.85	46.67± 0.00	54.69± 0.29	53.12± 0.36	71.58± 1.05	60.98± 0.62	57.84± 0.45
	0-shot MMLU	Vikhr 7B IT 0.4	69.15± 2.43	41.78± 0.89	52.09± 1.39	53.04± 1.06	77.89± 2.11	63.11± 1.45	57.60± 1.42

Table 2: Comparison of the results of the best models from each group (mean ± std)

smaller in the DSM dataset, as machine learning models in general can show low accuracy on limited data. On the other hand, if we refer to the study (Ignatiev et al., 2022), where less stringent restrictions on text volume and temporal proximity to questionnaire screening dates were applied to the same raw data, the results were still not very high: about 60% F1-macro for psycholinguistic features and TF-IDF features.

Secondly, the text format in the DSM introduces some noise. The texts in DSM are concatenated together with a collection of posts from users' personal pages, and they mostly lack coherent logic and cohesion in the resulting text. Even with the 6,000-character limit, the standard deviation in word count is approximately 300 words, compared with 100 in DE. Thus, DSM is a much noisier dataset, which can strongly affect the quality of classification with psycholinguistic markers, where the values of some markers can be affected by the volume of text analyzed, even considering their normalization with respect to the text volume.

Finally, the way in which the target class of pathology is defined can be of great importance. Although a widely used and validated psychological questionnaire was used for the DSM dataset, its results cannot be compared with a clinically confirmed diagnosis. The same findings can be outlined in work that criticizes the way in which social media users are labeled for mental illness by indirectly affiliating or self-identifying mental ill-health (Ernala et al., 2019). In favor of the significance of this factor is also the fact that TF-IDF-based features performed significantly better on DE data than on DSM, although, unlike psycholinguistic markers, they do not have an initial specialization for detecting signs of mental ill-health.

To take into account these considerations, we applied LLMs to this task and showed that the bigger model is able to partially overcome these issues.

Corpus	Mode	Model	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
	5-shot MMLU	Gemma2 9B IT	88.83± 0.25	96.56± 0.49	92.53± 0.36	59.56± 4.84	29.29± 1.43	39.26± 2.32	65.90± 1.34
	LoRA	Gemma2 2B IT	92.33± 0.94	97.14± 1.04	94.67± 0.72	76.44± 6.36	52.98± 6.33	**62.33± 5.60**	**78.50± 3.13**
	SFT	RuBERT	92.26± 0.52	96.22± 1.56	94.19± 0.68	72.12± 10.10	52.98± 3.81	60.60± 2.87	77.39± 1.72
D-all	0-shot	Gemma2 9B IT	93.28± 0.00	76.69± 0.00	84.18± 0.00	33.33± 0.00	67.86± 0.00	44.71± 0.00	64.44± 0.00
	0-shot MMLU	Gemma2 9B IT	94.87± 0.00	68.10± 0.00	79.29± 0.00	29.73± 0.00	78.57± 0.00	43.14± 0.00	61.21± 0.00
	SFT	Linguistic features	88.10± 0.60	88.10± 3.40	88.00± 1.60	51.60± 5.80	50.70± 4.10	50.80± 2.60	69.40± 1.90
	SFT	TF-IDF	89.50± 1.00	84.20± 1.00	86.80± 0.60	47.50± 1.80	59.10± 4.20	52.60± 2.50	69.70± 1.50
	5-shot MMLU	SaigaLlama3 8B	80.51± 5.64	11.26± 0.46	19.76± 0.88	48.46± 0.41	96.80± 1.07	**64.59± 0.60**	42.18± 0.74
	LoRA	SaigaLlama3 8B	58.56± 1.32	56.32± 4.83	57.33± 2.79	51.59± 2.03	53.78± 3.74	52.56± 1.83	54.95± 1.57
	SFT	RuRoBERTa	61.53± 3.69	60.15± 13.80	59.68± 4.47	54.39± 0.97	54.44± 18.08	52.27± 11.87	55.97± 3.88
	SFT	BERT	59.44± 1.74	46.93± 3.66	52.33± 2.10	50.43± 0.98	62.67± 4.81	55.82± 2.21	54.08± 1.08
A-all	0-shot	Qwen2 7B IT	60.01± 0.41	69.66± 0.92	64.47± 0.17	56.72± 0.10	46.13± 1.60	50.87± 0.99	57.67± 0.41
	0-shot	Gemma2 2B IT	71.44± 0.87	26.44± 0.00	38.59± 0.13	50.69± 0.15	87.73± 0.53	64.26± 0.27	51.42± 0.20
	0-shot MMLU	Qwen2 7B IT	62.67± 0.06	64.83± 1.38	63.72± 0.65	57.51± 0.47	55.20± 1.07	56.32± 0.32	**60.02± 0.17**
	SFT	Linguistic features	53.40± 2.90	39.70± 19.30	42.00± 17.90	46.10± 2.40	60.20± 20.40	51.00± 7.80	46.50± 6.20
	SFT	TF-IDF	51.80± 0.70	40.00± 15.60	43.30± 10.80	44.30± 2.10	56.40± 17.80	48.60± 8.50	46.00± 2.40

Table 3: Results of classification on D-all and A-all datasets, the best models from each group (mean ± std)

Transfer	Mode	Model	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
	LoRA	SaigaLlama3 8B	58.49± 3.68	83.95± 13.97	68.73± 7.39	40.34± 13.85	12.04± 4.99	17.21± 5.85	42.97± 4.20
	LoRA	Vikhr 7B IT 0.4	60.33± 5.19	61.73± 20.91	58.81± 10.42	37.58± 5.47	37.04± 24.57	33.98± 15.06	46.39± 4.16
	LoRA	Gemma2 2B IT	64.14± 3.85	71.60± 21.45	65.89± 10.13	51.07± 8.34	38.89± 22.91	38.67± 16.04	52.28± 6.44
	LoRA	Gemma2 9B IT	64.46± 5.24	66.67± 20.73	64.11± 11.57	51.24± 8.98	45.37± 18.54	44.36± 11.95	**54.23± 7.12**
	LoRA	Vikhr 7B IT 5.4	58.24± 5.42	67.28± 21.75	60.24± 11.51	31.33± 7.80	25.93± 24.36	25.09± 15.40	42.66± 4.75
	LoRA	Qwen2 7B IT	61.14± 3.36	61.11± 18.61	59.75± 11.42	43.62± 8.81	42.59± 15.93	41.25± 8.00	50.50± 5.32
DE to DSM test	LoRA	VikhrGemma 2B IT	60.84± 2.46	71.60± 16.93	64.90± 8.82	42.54± 7.31	31.48± 15.27	33.97± 12.07	49.44± 5.25
	SFT	RuRoBERTa	60.96± 7.01	41.98± 26.59	46.07± 14.36	33.75± 15.28	59.26± 27.34	42.95± 19.50	44.51± 5.07
	SFT	RuBioRoBERTa	61.65± 2.51	50.00± 23.88	52.54± 12.11	34.54± 15.55	52.78± 24.59	41.57± 18.71	47.05± 5.29
	SFT	RuBERT	68.00± 10.29	27.78± 9.01	38.68± 10.16	42.66± 2.80	80.56± 9.49	55.64± 4.19	47.16± 5.80
	SFT	BERT	76.85± 5.46	29.63± 8.28	42.21± 9.33	45.34± 2.29	87.04± 4.14	**59.54± 2.03**	50.88± 5.43
	SFT	Linguistic features	60.00± 2.50	57.40± 7.00	58.50± 4.50	40.00± 3.50	42.60± 6.90	41.10± 4.50	49.80± 3.10
	SFT	TF-IDF	59.90± 3.40	30.90± 9.20	40.00± 8.50	40.20± 1.50	69.40± 8.30	50.70± 2.70	45.40± 3.80

Table 4: Transfer models from DE to DSM (mean ± std)

While RuBERT and the model trained on linguistic features both show comparatively low results, a 5-shot LLM performs significantly better. The same outcomes are observed in the experiments with anxiety datasets.

For a better understanding of per-class performance of models, we also provided results for the best models on the DSM dataset for various splits within a class based on the initial BDI scores. The results are provided in Appendix F.

5.1.3 AL, AD, and AC Datasets

Turning to the anxiety detection task, the difference between the AutoML-based and encoder-based models is only noticeable on the picture description (AD) dataset, where the accuracy of the RuRoBERTa model was 55% F1-macro. Overall, it can be said that none of the non-LLM models performed well in this task. However, LLMs performed significantly better on these tasks. Models with LoRA, in general, perform better than encoder-based and AutoML-based models, but the best performance is achieved by 0-shot and 5-shot prompting. Again, this can be linked to the complexity of the domain and the small amount of training data, which makes it hard to fine-tune a model.

In general, the LLMs outperform other models on all datasets, but on the DE and DSM, the gap between various types of models is smaller, leaving the usage of non-LLM models reasonable for some specific cases.

5.2 Classification on D-all and A-all Datasets

We combined the depression and anxiety datasets into pooled datasets D-all (DE and DSM) and A-all (AL, AD, and AC). Classification performance of the best models on D-all and A-all combined data is presented in Table 3.

The LLMs performed better than other models on both D-all and A-all. It is also noticeable that methods with model tuning work better on D-all, while prompting shows better results on A-all. It can be explained with significant genre and length variations in A-all datasets, so the general-purpose models perform better than finetuned ones due to the complexity of the data.

5.3 Classification of DSM Data with Best Models from DE

The classification results of models that demonstrated the best performance on the DE dataset applied to DSM test data are shown in Table 4.

This experiment shows that models trained on essays from depressed subjects cannot be used to detect depression from social media texts from users who have taken a depression questionnaire and received a high score. Similar findings were shown in (Ernala et al., 2019), but with a reverse logic of the experiments.

Such results may also be just due to the fact that essay texts and collections of social network posts are very different in genre. A sample of social network users with clinically diagnosed depression would be needed to clarify this issue.

5.4 LLM Results Interpretation

As shown in Section 5.3, the overall best results were obtained on the DE dataset with LLMs. To further analyze the quality of these results from the clinical perspective, we conducted an additional experiment on LLM results interpretation.

We chose the best LLM in a 5-shot setting on the DE dataset (Vikhr 7B IT 5.4) and asked two psychologists with relevant experience to rate the LLM generation on a scale from 1 to 5. The LLM-generated answer in this setting consists of a predicted class label and a detailed explanation of why this label was chosen. Details of the rating scale and statistics are provided in Appendix D. The psychologists assessed only texts from patients with clinically diagnosed depression that the LLM correctly labeled.

The average score from expert psychologists for LLM explanations was assigned to 2.84 out of 5. Moreover, the psychologists also noted each explanation, which can be marked as true to some extent (e.g., with some true claims in the explanation) – there are approximately 66% of such explanations. These results show that the explanations for depression detection from texts contain both correct and erroneous parts. Overall, the LLM (in the described settings) does not generate enough explanations that meet the requirements of clinicians.

To further extend the error analysis in the LLM generations in the psychological domain, we asked psychologists to describe the most common errors in the generated explanations. To do so, psychologists annotated each explanation with the list of errors and categorized these mistakes into the following groups: (1) tautology, (2) groundless generalization, (3) false conclusion, (4) confabulation, (5) distortion of medical understanding of depression, (6) incompleteness of selected signs of depression. Description and examples of these error types are located in Table 17 in the Appendix D. It should also be noted that from the general NLP perspective, most of these types of errors can be defined as hallucinations; however, we suppose that a more precise definition is needed in this specific domain.

On average, each explanation contains two or more errors. The most common types of errors are groundless generalization, false conclusion, and confabulation, which occur in 56%, 56%, and 50% of samples, respectively. However, these errors appear several times in each explanation. The incompleteness of selected signs of depression appears in 44%, while distortion of medical understanding of depression and tautology are the rare errors, which appear in 19% and 13%. These results show that there exist various types of errors, specific to the mental health disorders domain. The additional results with detailed analysis of feature importance for the pathology and healthy class prediction are located in Appendix E.

6 Conclusion

In this paper, we have investigated the effectiveness of traditional machine learning methods and LLMs on the task of detecting depression and anxiety. The results obtained in our work establish the new state-of-the-art on five Russian-language datasets.

Our investigation shows that psycholinguistic features can produce results at the level of encoder-based models when texts from individuals with clinically diagnosed depression are used for training. BERT models, in turn, perform better on noisy text data, where examples from the training sample may vary widely in text length or genre. LLM-based models performed best on all five different mental health datasets. Even without fine-tuning, LLMs usually demonstrate relatively high performance. In response to **RQ1**, the experimental results indicate that LLM-based models have high potential for detecting mental disorders from texts.

In response to **RQ2**, the findings reveal that models trained on essays from depressed individuals are not effective for detecting depression in social media texts from users who have completed a de-

pression questionnaire and scored high.

Finally, we evaluated LLM-generated explanations and showed that in the current state, these explanations do not meet the clinicians' requirements with an overall score of 2.84 out of 5, which answers **RQ3**. We also constructed a detailed classification of common errors in LLM explanations from the clinicians' perspective, to guide further improvements of LLMs in this domain.

Limitations

The main limitation of this paper is that it is impossible to share all the raw texts from the used datasets, as they are all distributed under different terms. However, we present several anonymized textual examples and extracted psycholinguistic features.

The amount of sample data used for predicting anxiety is small, which does not allow us to adequately judge the possibility of predicting anxiety in Russian text using machine learning methods. Although the results on all the anxiety datasets used show poor accuracy, perhaps a very different scale of data is needed for this task. The paper does not discuss the differences between the results within one group of the models in detail, as this was not the aim of the paper.

We conducted the experiments only for Russian due to the poor availability of related datasets for other languages. However, the used methods in general are language agnostic, so the results could be extended to other languages. The overall result about LLMs as the best-performing models matches similar studies for English on closed data.

The final limitation is the temporal validity of the results. Due to the fast growth of LLM-based solutions, future LLMs could outperform the obtained results. However, we suppose that the current state of the field already presents an interest and is therefore investigated in the paper.

Ethical Considerations

The problem discussed in this paper is the sensitive issue of mental health. To avoid any possible harm, we did not fully open-source the used datasets. All shared data does not contain any information that names or uniquely identifies individuals, nor does it contain offensive content. The examined models for the detection of mental disorders do not aim to replace a professional physician; on the contrary, these models are intended to support a human expert.

Acknowledgements

The work of Gleb Kuzmin was supported by a grant, provided by the MED of Russia (agreement identifier 000000C313925P4G0002) and the agreement with the ISP RAS dated June 20, 2025 No. 139-15-2025-011.

The research was carried out using the infrastructure of the shared research facilities «High Performance Computing and Big Data » of FRC CSC RAS (CKP «Informatics»)

References

Luna Ansari, Shaoxiong Ji, Qian Chen, and Erik Cambria. 2023. Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on Computational Social Systems*, 10:211–219.

Nirmal Varghese Babu and Edward Grace Mary Kanaga. 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *SN Computer Science*, 3(1):74.

Aaron T. Beck, Calvin H. Ward, Mock Mendelson, Jeremiah Mock, and John Erbaugh. 1961. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571.

Ingvar Bjelland, Alv A Dahl, Tone Tangen Haug, and Dag Neckelmann. 2002. The validity of the hospital anxiety and depression scale: an updated literature review. *Journal of psychosomatic research*, 52(2):69–77.

Iacer Calixto, Victoria Yaneva, and Raphael Moura Cardoso. 2022. Natural language processing for mental disorders: an overview. *Natural Language Processing In Healthcare*, pages 37–59.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Leonard R. Derogatis. 1983. SCL-90-R: Administration, scoring and procedures. *Manual II for the R (evised) Version and Other Instruments of the Psychopathology Rating Scale Series*.

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.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 134. ACM.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2962–2970.

Muskan Garg. 2023. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842.

Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21:1–18.

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, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Sharath Chandra Guntuku, David B. Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Bakir Hadzic, Parvez Mohammed, Michael Danner, Julia Ohse, Yihong Zhang, Youssef Shiban, and Matthias Rätsch. 2024. Enhancing early depression detection with ai: a comparative use of nlp models. *SICE Journal of Control, Measurement, and System Integration*, 17(1):135–143.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Nikolay Ignatiev, Ivan V. Smirnov, and Maxim Stankevich. 2022. Predicting depression with text, image, and profile data from social media. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2022, Online Streaming, February 3-5, 2022*, pages 753–760. SCITEPRESS.

Jinwoo Jeong, Sujin Yoon, Dongyoung Sohn, and Yong Suk Choi. 2023. Reading emotions in the digital age: A deep learning approach to detecting anxiety during the covid-19 pandemic through social media. *International Journal of Communication*, 17:22.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, and Yong Li. 2024. Depression detection on social media with large language models. *arXiv preprint arXiv:2403.10750*.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Tatiana Litvinova and Ekarerina Ryzhkova. 2018. Rusneuropsych: open corpus for study relations between author demographic, personality traits, lateral preferences and affect in text. *International Journal of Open Information Technologies*, 6(3):32–36.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, LA. Association for Computational Linguistics.

Tobias Mayer, Neha Warikoo, Amir Eliassaf, Dana Atzil-Slonim, and Iryna Gurevych. 2024. Predicting

client emotions and therapist interventions in psychotherapy dialogues. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1463–1477, St. Julian's, Malta. Association for Computational Linguistics.

Tatyana I. Medvedeva, Sergei N. Enikolopov, Olga M. Boyko, Oksana Yu. Vorontsova, and Maxim A. Stankevich. 2021. Lexical analysis of statements about covid-19 of people with a high level of somatization. *Lomonosov Psychology Journal*, 14(3):39–64.

Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, December 13-16, 2016*, pages 136–143. IEEE.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for Russian. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 189–199, Miami, Florida, USA. Association for Computational Linguistics.

Mahmud Omar and Inbar Levkovich. 2025. Exploring the efficacy and potential of large language models for depression: A systematic review. *Journal of Affective Disorders*, 371:234–244.

David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. 2020. Towards preemptive detection of depression and anxiety in Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 82–89, Barcelona, Spain (Online). Association for Computational Linguistics.

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. *arXiv preprint arXiv:2305.05138*.

Fabien Ringeval, Björn W. Schuller, Michel F. Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, October 23 - 27, 2017*, pages 3–9. ACM.

Faisal Muhammad Shah, Farzad Ahmed, Sajib Kumar Saha Joy, Sifat Ahmed, Samir Sadek, Rimon Shil, and Md. Hasanul Kabir. 2020. Early depression detection from social network using deep learning techniques. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 823–826.

Ivan V. Smirnov, Maksim Stankevich, Yulia Kuznetsova, Margarita Suvorova, Daniil Larionov, Elena Nikitina, Mikhail Savelov, and Oleg G. Grigoriev. 2021. TITANIS: A tool for intelligent text analysis in social media. In *Artificial Intelligence - 19th Russian Conference, RCAI 2021, Taganrog, Russia, October 11-16, 2021, Proceedings*, pages 232–247. Springer.

Maksim Stankevich, Yulia Kuznetsova, Ivan Smirnov, Natalia Kiselnikova, and Sergey Enikolopov. 2019. Predicting depression from essays in russian. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 647–657.

Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

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, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Vankayala Tejaswini, Korra Sathya Babu, and Bibhudatta Sahoo. 2024. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 23(1):4:1–4:20.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126, St. Julians, Malta. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alexander Yalunin, Alexander Nesterov, and Dmitriy Umerenkov. 2022. Rubioroberta: a pre-trained biomedical language model for russian language biomedical text mining. *arXiv preprint arXiv:2204.03951*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan

Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tianlin Zhang, Annika Marie Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

A Hyperparameters

The used net and optimal parameters, alongside the used checkpoint of encoder models, are presented in Table 7. For the models trained on the psycholinguistic features and TF-IDF, the best model was selected with the AutoML pipeline, with the most models chosen as a Random Forest ensemble of up to four models.

The hyperparameters for LLMs, tuned with LoRA, are presented in Table 8. The LoRA was applied to the following target layers: `q_proj`, `up_proj`, `o_proj`, `k_proj`, `down_proj`, `gate_proj`, `v_proj`. For LLMs 0-shot and 5-shot evaluation, we used the following generation parameters: `do_sample=False`, `temperature=1.0`, `top_p=1.0`, `top_k=50`, `repetition_penalty=1.0`. The predicted classes were extracted using the string matching algorithm. For MMLU-style evaluation, the generation length was limited to one token, while for regular evaluation, it was set to 64 tokens. Prompt examples are presented in Table 5. For the 0-shot and 5-shot evaluation, we averaged the results on several prompts with various system parts to take into account the sensitivity of the generation with respect to the prompt. These prompt variations are presented in Table 6. In all 0-shot and 5-shot experiments, we did not conduct best-of-N aggregation of answers and used one extracted answer per generation, which was averaged over several prompts to ensure generalizability of results.

The used hardware, as well as GPU hours, carbon footprint, and memory requirements are presented in Table 9.

B Full Results for Models

Table 10 shows results for TF-IDF, models on linguistic features, and encoder-based transformers. Tables 11 to 13 contain results for LLMs with PEFT, 0-shot and 5-shot evaluation, and 0-shot and 5-shot MMLU-style evaluation correspondingly.

C Full Results on D-all and A-all Datasets

Tables 14 to 16 show results for all models on D-all and A-all datasets. Among models with SFT Gemma2 2B IT is the best on D-all, while on A-all the best are encoder-based models – RuRoBERTa and BERT. While in general prompting performs worse than SFT on D-all, the best model (Gemma 2 9B IT) shows comparable performance with AutoML models. On the contrary, prompting on the A-all shows better results than SFT. The best models in various prompting settings are Qwen2 7B IT, SaigaLlama3 8B, and Gemma2 2B IT.

D Psychologists Evaluation Setup

During the LLM generation interpretation, psychologists used the following rating scale for explanations, given by LLM: 1 - completely erroneous, 2 - mostly erroneous, 3 - partially erroneous, partially correct, 4 - mostly correct, 5 - completely correct. On this scale, the psychologists rated LLM explanations as 2.84 out of 5 with Fleiss' $\kappa = 0.39$, which shows fair inter-annotator agreement. To better categorize errors in LLM explanations, psychologists also provided a list of the most common errors in each generation with detailed explanations of the error type. This list is provided in Table 17.

We also conducted an additional experiment with a clinically informed prompt to explore how the additional information from psychologists can affect the quality of explanations. To do so, we asked psychologists to select parts of the text that can serve as depression markers for texts that were used in a 5-shot prompt for the DE dataset. After we performed a 5-shot evaluation of the best model with a modified 5-shot prompt, which now includes depression markers along with the target label. The results are presented in Table 18. The adapted prompt leads to a slightly better F1-macro score; however, such a prompt requires qualified psychologists to label several texts for each investigated dataset. These results provide an interesting direction for future work and show that the overall quality of LLMs in mental disorders detection tasks can be improved with the help of psychologists.

To further investigate the quality of model explanations with clinically informed prompts, we asked psychologists to conduct the evaluation of results as in Section 5.4. During evaluation, the psychologists

Task	Mode	Prompt example (system & user)
Depression	0-shot	*System:* You play the role of a psychologist's assistant who helps diagnose the presence or absence of a depressive disorder. You will be given a text written by a person. Determine the author's depression level from the text, where 0 is no depression, 1 is depression, and then write why you chose this answer. *User:* Text: {input_text} Answer (0 or 1):
Depression	5-shot	*System:* You play the role of a psychologist's assistant who helps diagnose the presence or absence of a depressive disorder. You will be given a text written by a person. Determine the author's depression level from the text, where 0 is no depression, 1 is depression, and then write why you chose this answer. *User:* Text: {example_1} Answer (0 or 1): 0 Text: {example_2} Answer (0 or 1): 0 Text: {example_3} Answer (0 or 1): 0 Text: {example_4} Answer (0 or 1): 1 Text: {example_5} Answer (0 or 1): 1 Text: {input_text} Answer (0 or 1):
Anxiety	0-shot	*System:* You play the role of a psychologist's assistant who helps diagnose the presence or absence of an anxiety disorder. You will be given a text written by a person. Determine the level of anxiety of the author of the text, where 0 is no anxiety, 1 is anxiety, and then write why you chose this answer. *User:* Text: {input_text} Answer (0 or 1):
Anxiety	5-shot	*System:* You play the role of a psychologist's assistant who helps diagnose the presence or absence of an anxiety disorder. You will be given a text written by a person. Determine the level of anxiety of the author of the text, where 0 is no anxiety, 1 is anxiety, and then write why you chose this answer. *User:* Text: {example_1} Answer (0 or 1): 0 Text: {example_2} Answer (0 or 1): 0 Text: {example_3} Answer (0 or 1): 0 Text: {example_4} Answer (0 or 1): 1 Text: {example_5} Answer (0 or 1): 1 Text: {input_text} Answer (0 or 1):

Table 5: Prompts for LLMs evaluation. For a better understanding, we present a translated English version (originally we used the same prompts in Russian). For models without system prompts (e. g. Gemma2) we concatenated the system prompt to the user prompt. Text in italics was used to denote system and user roles and was replaced by model-specific templates during evaluation.

Task	Prompt number	Prompt example (only system)
Depression	1	*System:* Read the provided text and determine whether the author has signs of depression. Use the scale: 0 - no depression, 1 - depression. Then explain why you chose this option.
Depression	2	*System:* Evaluate the text for the author's depressive state. Scale: 0 - no depression, 1 - depression is present. Justify your choice after indicating the answer.
Depression	3	*System:* Analyze this text and identify the presence of depressive manifestations in its author. Use a binary assessment: 0 - no depression, 1 - presence of depression. Provide a rationale for your decision.
Depression	4	*System:* Assess the psychological state of the author of the text in the context of depression. Use the gradation: 0 - no signs of depression, 1 - there are signs of depression. Detail the reasons for making your decision.
Anxiety	1	*System:* Read the provided text and determine whether the author has signs of anxiety. Use the scale: 0 - no anxiety, 1 - anxiety. Then explain why you chose this option.
Anxiety	2	*System:* Evaluate the text for the author's anxiety. Scale: 0 - no anxiety, 1 - anxiety is present. Justify your choice after indicating the answer.
Anxiety	3	*System:* Analyze this text and identify the presence of anxiety in its author. Use a binary assessment: 0 - no anxiety, 1 - anxiety present. Justify your decision.
Anxiety	4	*System:* Assess the psychological state of the author of the text in the context of anxiety. Use the gradation: 0 - no signs of anxiety, 1 - signs of anxiety present. Detail the reasons for making your decision.

Table 6: Additional prompts used to average results for LLMs evaluation. Here we present only the examples of system prompts, because the other prompt parts remained unchanged as in Table 5.

Model name	Corpus	Num. of epochs	Learning rate	Batch size	Weight decay
RuBERT	A-all	15	5e-05	16	0.10
	D-all	15	5e-05	16	0.10
	DE	10	7e-05	32	0.10
	AD	2	6e-06	4	0.10
	AL	2	9e-06	4	0.10
	AC	2	6e-06	4	0.10
	DSM	13	1e-04	4	0.00
RuRoBERTa	A-all	13	9e-06	16	0.01
	D-all	11	6e-06	8	0.00
	DE	9	9e-06	4	0.00
	AD	7	2e-05	4	0.00
	AL	11	6e-06	8	0.00
	AC	6	1e-04	32	0.01
	DSM	13	5e-06	16	0.10
RuBioRoBERTa	A-all	15	3e-05	16	0.01
	D-all	11	7e-06	32	0.00
	DE	8	1e-05	4	0.00
	AD	2	7e-06	8	0.10
	AL	9	2e-05	16	0.01
	AC	8	1e-05	4	0.00
	DSM	15	9e-06	16	0.01
BERT	A-all	14	7e-06	16	0.01
	D-all	15	5e-05	16	0.10
	DE	15	5e-05	16	0.10
	AD	2	6e-06	4	0.10
	AL	2	6e-06	4	0.10
	AC	4	1e-04	32	0.01
	DSM	5	1e-04	32	0.01

Table 7: Optimal hyperparameters for transformer models. We used the following net for hyperparameters tuning: learning rate - [5e-6, 6e-6, 7e-6, 9e-6, 1e-5, 2e-5, 3e-5, 5e-5, 7e-5, 1e-4], num. of epochs - from 2 to 15, batch size - [4, 8, 16, 32], weight decay - [0, 0.01, 0.1].

Model name	Corpus	Num. of epochs	Learning rate	Batch size	Weight decay	LoRA α	LoRA dropout	LoRA rank
SaigaLlama3 8B	DE	13	1e-04	4	1e-01	16	1e-01	8
	DSM	3	1e-05	16	1e-02	32	5e-02	16
	AL	15	1e-04	4	1e-01	16	1e-01	16
	AD	15	1e-04	4	1e-01	16	1e-01	16
	AC	11	7e-05	16	1e-01	32	5e-02	8
	K-all	14	3e-05	16	0e+00	16	5e-02	16
	D-all	11	7e-05	16	1e-01	32	5e-02	8
Vikhr 7B IT 0.4	DE	2	5e-05	4	1e-01	32	5e-02	8
	DSM	11	7e-05	16	1e-01	32	5e-02	8
	AL	4	9e-06	8	1e-02	16	1e-01	16
	AD	11	1e-04	4	0e+00	32	1e-01	16
	AC	3	1e-04	16	0e+00	16	5e-02	16
	K-all	11	2e-05	16	1e-02	16	1e-01	16
	D-all	11	7e-05	16	1e-01	32	5e-02	8
Vikhr 7B IT 5.4	DE	15	1e-04	4	1e-01	16	1e-01	16
	DSM	4	9e-06	8	1e-02	16	1e-01	16
	AL	7	6e-06	8	0e+00	16	1e-01	8
	AD	8	5e-06	4	1e-02	16	1e-01	16
	AC	15	1e-04	4	0e+00	16	1e-01	8
	K-all	12	3e-05	16	0e+00	16	1e-01	8
	D-all	8	5e-06	4	1e-02	16	1e-01	16
VikhrGemma 2B IT	DE	15	1e-04	4	0e+00	16	1e-01	8
	DSM	2	6e-06	8	0e+00	16	1e-01	16
	AL	5	5e-05	4	1e-02	16	5e-02	16
	AD	8	2e-05	4	0e+00	32	5e-02	16
	AC	2	5e-05	4	1e-01	32	5e-02	8
	K-all	11	2e-05	16	1e-02	16	1e-01	16
	D-all	15	1e-04	4	1e-01	16	1e-01	16
Gemma2 2B IT	DE	15	1e-04	4	1e-01	16	1e-01	16
	DSM	8	7e-05	16	0e+00	32	5e-02	8
	AL	13	1e-04	4	0e+00	16	1e-01	8
	AD	12	7e-05	16	0e+00	32	5e-02	8
	AC	14	1e-04	4	0e+00	32	1e-01	16
	K-all	11	7e-05	16	1e-01	32	5e-02	8
	D-all	15	1e-04	4	1e-01	16	1e-01	16
Gemma2 9B IT	DE	9	1e-04	8	0e+00	16	5e-02	8
	DSM	13	5e-06	8	1e-02	16	1e-01	16
	AL	6	1e-04	8	0e+00	16	1e-01	16
	AD	11	2e-05	16	1e-02	16	1e-01	16
	AC	15	1e-04	4	0e+00	16	1e-01	8
	K-all	11	7e-05	16	1e-01	32	5e-02	8
	D-all	15	1e-04	4	0e+00	16	1e-01	8
Qwen2 7B IT	DE	5	5e-05	4	1e-02	16	5e-02	16
	DSM	11	7e-05	16	1e-01	32	5e-02	8
	AL	4	5e-05	8	0e+00	16	5e-02	8
	AD	11	7e-05	16	1e-01	32	5e-02	8
	AC	9	3e-05	4	0e+00	32	1e-01	8
	K-all	11	7e-05	16	1e-01	32	5e-02	8
	D-all	13	1e-04	8	1e-01	32	5e-02	8

Table 8: Optimal hyperparameters for LLMs with LoRA. We used the following net for hyperparameters tuning: learning rate - [5e-6, 6e-6, 7e-6, 9e-6, 1e-5, 2e-5, 3e-5, 5e-5, 7e-5, 1e-4], num. of epochs - from 2 to 15, batch size - [4, 8, 16, 32], weight decay - [0, 0.01, 0.1], LoRA α - [16, 32], LoRA dropout - [0.05, 0.1], LoRA rank - [8,16].

GPU type	NVIDIA V100 32GB	NVIDIA A100 80 GB	NVIDIA H100 80 GB	Total
GPU Hours	119	371	43	533
Carbon footprint, kg CO_2	11.50	29.87	3.46	44.83

Table 9: The approximate number of GPU hours and carbon footprint for all experiments.

Corpus	Model	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
DE	RuRoBERTa	90.44± 4.62	95.93± 2.28	92.98± 1.92	64.70± 29.29	56.82± 25.81	60.37± 27.21	76.67± 14.52
	RuBioRoBERTa	91.92± 5.21	95.19± 3.05	93.36± 2.13	64.25± 29.33	63.64± 28.63	63.73± 28.62	78.54± 15.30
	RuBERT	94.45± 1.16	91.11± 1.70	92.74± 1.04	68.41± 4.04	78.03± 4.85	72.80± 3.43	82.77± 2.21
	BERT	92.97± 0.80	92.96± 1.23	92.96± 0.85	71.34± 3.96	71.21± 3.39	71.23± 3.14	82.09± 1.98
	Linguistic features	94.00± 0.80	95.20± 1.00	94.60± 0.80	79.30± 4.00	75.00± 3.50	**77.00± 3.30**	**85.80± 2.10**
	TF-IDF	91.60± 2.10	94.40± 2.20	93.00± 1.20	74.80± 7.00	64.40± 10.30	68.50± 6.60	80.70± 3.80
DSM	RuRoBERTa	61.31± 4.38	73.46± 12.73	66.58± 7.66	46.76± 10.29	31.48± 6.93	36.68± 6.20	51.63± 5.80
	RuBioRoBERTa	62.10± 2.98	62.96± 6.05	62.46± 4.30	43.66± 4.43	42.59± 4.14	43.01± 3.70	52.74± 3.67
	RuBERT	60.78± 1.75	96.91± 6.90	74.52± 1.07	9.09± 20.33	5.56± 12.42	6.90± 15.42	40.71± 7.18
	BERT	61.09± 1.90	69.14± 14.61	64.13± 5.24	34.59± 15.71	33.33± 16.97	33.71± 16.01	48.92± 5.62
	Linguistic features	62.80± 1.80	59.30± 7.40	60.70± 4.00	43.70± 2.70	47.20± 7.70	45.10± 3.80	**52.90± 2.00**
	TF-IDF	62.20± 2.60	51.20± 13.90	55.30± 8.50	42.90± 3.60	53.70± 11.40	**46.90± 4.70**	51.10± 3.50
AL	RuRoBERTa	56.82± 5.65	75.00± 7.76	64.49± 5.59	52.54± 10.76	33.33± 12.77	40.17± 12.48	52.33± 8.31
	RuBioRoBERTa	53.62± 1.09	78.79± 15.67	63.22± 5.04	30.66± 21.79	21.05± 15.79	24.76± 18.06	43.99± 6.71
	RuBERT	53.63± 3.20	77.27± 19.99	62.08± 7.52	45.83± 29.56	21.93± 20.48	24.65± 17.57	43.36± 5.79
	BERT	52.71± 1.02	93.94± 8.16	67.44± 2.96	6.25± 13.98	2.63± 5.88	3.70± 8.28	35.57± 2.99
	Linguistic features	47.40± 21.40	36.40± 17.80	40.90± 19.20	48.50± 2.70	68.40± 14.60	56.10± 3.70	48.50± 8.20
	TF-IDF	61.50± 5.00	46.20± 3.10	52.60± 2.30	51.20± 2.50	65.80± 7.90	**57.50± 4.50**	**55.00± 2.90**
AD	RuRoBERTa	57.63± 2.52	56.67± 4.71	57.07± 3.19	52.80± 2.82	53.70± 4.14	**53.17± 2.84**	**55.12± 2.58**
	RuBioRoBERTa	45.06± 5.52	43.33± 19.72	41.49± 13.80	38.29± 8.58	40.74± 23.28	37.72± 12.46	39.60± 5.38
	RuBERT	48.63± 7.83	80.00± 29.58	59.34± 15.36	12.59± 18.14	12.96± 18.33	12.70± 18.04	36.02± 4.94
	BERT	52.03± 1.00	93.33± 7.99	66.74± 2.88	20.56± 21.12	4.63± 4.99	7.34± 7.71	37.04± 2.63
	Linguistic features	50.80± 3.30	43.30± 12.50	45.30± 8.10	44.70± 2.60	51.90± 15.90	47.30± 7.40	46.30± 1.80
	TF-IDF	54.90± 15.10	37.50± 4.80	43.20± 2.50	44.60± 7.80	59.30± 21.00	50.40± 12.70	46.80± 7.10
AC	RuRoBERTa	45.70± 20.48	77.04± 35.48	57.19± 25.66	24.57± 24.96	24.56± 35.49	21.22± 23.82	39.20± 7.55
	RuBioRoBERTa	52.92± 1.41	76.30± 14.15	62.09± 5.54	34.88± 15.72	20.18± 11.64	24.84± 12.88	43.47± 4.21
	RuBERT	54.29± 1.57	87.04± 19.35	66.00± 8.24	43.59± 22.45	14.04± 17.10	17.15± 15.70	41.58± 4.88
	BERT	57.01± 5.15	62.96± 16.81	58.44± 6.53	46.07± 8.77	41.67± 22.14	41.28± 17.15	49.86± 6.84
	Linguistic features	64.80± 16.00	45.60± 21.00	47.00± 19.60	48.70± 3.60	60.50± 20.10	**52.60± 7.30**	49.80± 7.80
	TF-IDF	56.30± 2.10	63.30± 4.20	59.50± 2.50	48.90± 3.30	41.70± 5.60	44.90± 4.30	**52.20± 2.70**

Table 10: The results for encoder models and AutoML models on the five main datasets.

set the average score to 2.47 out of 5, which is lower than for prompting without a clinically informed prompt. Due to the significant difference in the explanations of LLM, the mistakes were categorized into other groups, namely: (1) confabulation, (2) undifferentiation (inability to select symptomatic parts of the text), and (3) incompleteness of explanation. With a clinically informed prompt, each of the explanations contains at least one of the mentioned mistakes. The most common types of mistakes are confabulation and undifferentiation, which are found in 47% of the explanations. The incompleteness of the explanation occurs in 40% of the explanations. Overall, these results show that even with the clinically informed prompt, modern LLMs are unable to generate explanations that meet the requirements of clinicians.

E Feature Importance Ablation

To further deepen the explanation analysis, we conducted feature ablation for the best LLM on the DE dataset. For this purpose, we extracted the top-3 words by their importance on target label generation (and on full explanation generation) for each text in the test set. The words were extracted with the Feature Ablation method from the Captum framework (Kokhlikyan et al., 2020), which calculates feature ablation based on differences in predictions with and without features. The list of unique words with the biggest mean feature importance is shown at Figure 1 (importance for target label generation) and Figure 2 (importance for full explanation generation). For both target class and explanation generation, the most important features in texts with predicted pathology class contain a significant amount of words with negative meanings (such as "disappointing", "negative", "bitter", "angry", "disgusting", "darkness", etc.), as well as words with direct pathology description ("depression", "fear"). The detailed analysis from the trained clinicians reveals that most of the features with high importance for depression class are connected with fear, suffering, and unhealthy conditions. On the other hand, the most important features of a healthy class contain mostly positive semantics, such as "humanity", "unselfishness", and "kind".

Corpus	Model	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
DE	Gemma2 2B IT	92.96± 1.39	97.22± 1.40	95.02± 0.48	86.63± 4.79	69.70± 6.78	76.84± 3.12	85.93± 1.75
	Gemma2 9B IT	91.90± 0.90	96.48± 1.00	94.13± 0.58	82.11± 3.76	65.15± 4.29	72.53± 2.91	83.33± 1.72
	Qwen2 7B IT	90.18± 0.34	95.19± 2.10	92.60± 0.96	75.46± 7.33	57.58± 2.14	65.05± 2.19	78.83± 1.56
	SaigaLlama3 8B	91.59± 0.79	98.70± 1.19	95.01± 0.41	92.88± 5.97	62.88± 4.08	74.74± 2.12	84.87± 1.22
	Vikhr 7B IT 0.4	90.26± 1.15	94.07± 2.77	92.10± 1.11	72.82± 12.43	58.33± 6.11	63.92± 3.46	78.01± 2.17
	Vikhr 7B IT 5.4	90.80± 2.36	95.74± 2.52	93.15± 1.15	79.07± 9.93	59.85± 12.14	66.91± 8.35	80.03± 4.65
	VikhrGemma 2B IT	94.74± 0.88	96.48± 1.62	95.59± 0.66	84.96± 5.72	78.03± 4.08	**81.13± 2.42**	**88.36± 1.52**
DSM	Gemma2 2B IT	63.29± 4.98	58.64± 9.18	60.62± 6.33	44.42± 5.41	49.07± 9.31	46.24± 6.49	53.43± 5.33
	Gemma2 9B IT	64.20± 3.12	56.79± 6.65	60.16± 4.97	45.12± 4.14	52.78± 4.24	48.54± 3.57	54.35± 3.90
	Qwen2 7B IT	68.87± 6.39	72.22± 8.21	70.40± 6.75	55.33± 11.08	50.93± 10.84	**52.81± 10.26**	**61.61± 8.32**
	SaigaLlama3 8B	57.46± 7.16	56.17± 15.48	56.13± 10.59	37.43± 9.05	38.89± 13.98	37.21± 10.11	46.67± 7.70
	Vikhr 7B IT 0.4	64.12± 2.52	70.99± 11.40	67.06± 5.90	50.75± 11.53	40.74± 7.64	44.07± 4.72	55.57± 4.06
	Vikhr 7B IT 5.4	55.50± 5.73	54.32± 8.73	54.79± 6.93	34.30± 7.65	35.19± 7.64	34.58± 7.06	44.69± 6.59
	VikhrGemma 2B IT	61.41± 6.24	58.64± 25.16	57.67± 18.04	48.66± 12.71	49.07± 16.79	45.68± 6.56	51.68± 8.61
AL	Gemma2 2B IT	60.18± 4.51	60.61± 13.55	59.66± 8.17	54.87± 7.30	53.51± 10.71	**53.45± 6.06**	**56.56± 5.40**
	Gemma2 9B IT	47.52± 5.48	46.21± 4.08	46.79± 4.42	39.00± 6.04	40.35± 8.95	39.60± 7.34	43.20± 5.65
	Qwen2 7B IT	53.25± 10.65	50.00± 13.38	51.40± 11.90	46.74± 11.99	50.00± 11.67	48.16± 11.50	49.78± 11.33
	SaigaLlama3 8B	56.45± 3.35	53.03± 6.25	54.32± 3.10	48.33± 4.02	51.75± 11.13	49.67± 7.13	51.99± 3.55
	Vikhr 7B IT 0.4	54.87± 5.68	50.00± 6.94	52.04± 4.93	47.01± 4.37	51.75± 10.27	49.01± 6.60	50.53± 4.71
	Vikhr 7B IT 5.4	57.33± 3.30	56.06± 8.16	56.47± 4.96	50.76± 4.71	51.75± 6.39	50.99± 3.89	53.73± 3.56
	VikhrGemma 2B IT	59.30± 2.53	56.82± 8.60	57.59± 4.70	52.10± 3.11	54.39± 9.92	52.81± 5.51	55.20± 2.95
AD	Gemma2 2B IT	59.21± 5.93	55.83± 10.17	56.62± 6.27	52.58± 5.69	55.56± 15.38	53.36± 9.71	54.99± 5.86
	Gemma2 9B IT	52.03± 4.74	46.67± 5.53	49.05± 4.48	46.50± 4.47	51.85± 8.28	48.91± 5.93	48.98± 4.50
	Qwen2 7B IT	48.10± 2.19	49.17± 12.72	47.91± 7.44	41.98± 4.13	41.67± 13.13	41.08± 7.91	44.50± 2.76
	SaigaLlama3 8B	57.84± 6.15	53.33± 4.71	55.22± 3.94	51.13± 5.98	55.56± 12.42	53.01± 8.89	54.12± 5.88
	Vikhr 7B IT 0.4	59.81± 4.42	51.67± 12.13	54.66± 7.59	53.47± 4.23	61.11± 11.56	**56.43± 6.19**	**55.55± 4.55**
	Vikhr 7B IT 5.4	51.91± 8.89	55.00± 12.58	53.32± 10.58	47.66± 10.86	44.44± 8.49	45.88± 9.46	49.60± 9.75
	VikhrGemma 2B IT	45.93± 4.26	39.17± 6.72	42.13± 5.50	42.09± 3.57	49.07± 5.93	45.22± 4.01	43.68± 3.91
AC	Gemma2 2B IT	58.73± 3.32	63.70± 4.19	61.08± 3.43	52.17± 4.36	46.93± 5.35	49.35± 4.71	55.22± 3.84
	Gemma2 9B IT	56.84± 1.85	61.48± 5.97	58.85± 2.17	49.07± 2.00	44.30± 8.78	46.21± 5.61	52.53± 2.19
	Qwen2 7B IT	57.36± 3.77	57.04± 5.97	57.15± 4.76	49.73± 4.74	50.00± 4.02	49.81± 4.08	53.48± 4.25
	SaigaLlama3 8B	59.02± 3.26	65.56± 5.70	62.03± 3.81	53.16± 4.52	46.05± 5.84	49.19± 4.61	55.61± 3.74
	Vikhr 7B IT 0.4	59.62± 3.49	60.37± 8.36	59.78± 5.60	52.72± 4.82	51.75± 6.56	52.00± 4.61	55.89± 4.23
	Vikhr 7B IT 5.4	61.87± 2.07	62.96± 8.18	62.13± 4.30	55.50± 3.03	53.95± 7.25	**54.33± 3.58**	**58.23± 2.38**
	VikhrGemma 2B IT	57.77± 2.90	64.81± 6.21	60.98± 3.77	51.40± 4.47	43.86± 6.39	47.11± 4.91	54.05± 3.57

Table 11: The results for LLMs with LoRA on the five main datasets.

F Disaggregated Results for DSM Dataset

The scores for the Depression-Social Media (DSM) dataset were aggregated based on the obtained BDI questionnaire scores. To further investigate how these results vary between subgroups with various scores within one group (inside sub-splits of healthy or pathology class), we provided Tables 19 and 20 with disaggregated scores. As one can see, for the healthy class, the F1-healthy scores are mostly consistent for all models, with the exception of RuBioRoBERTa, which has a significantly lower F1-healthy score for the split with BDI Score from 3 to 6. For pathology class, F1-pathology scores vary more significantly, which can be described with a bigger variability of initial scores, especially for the split with scores from 34 to 63.

G Textual Examples

For a better understanding of the investigated task, we presented several samples from the DE dataset - one per class. The samples were paraphrased and anonymized, and then translated into English. The examples are presented in Table 21.

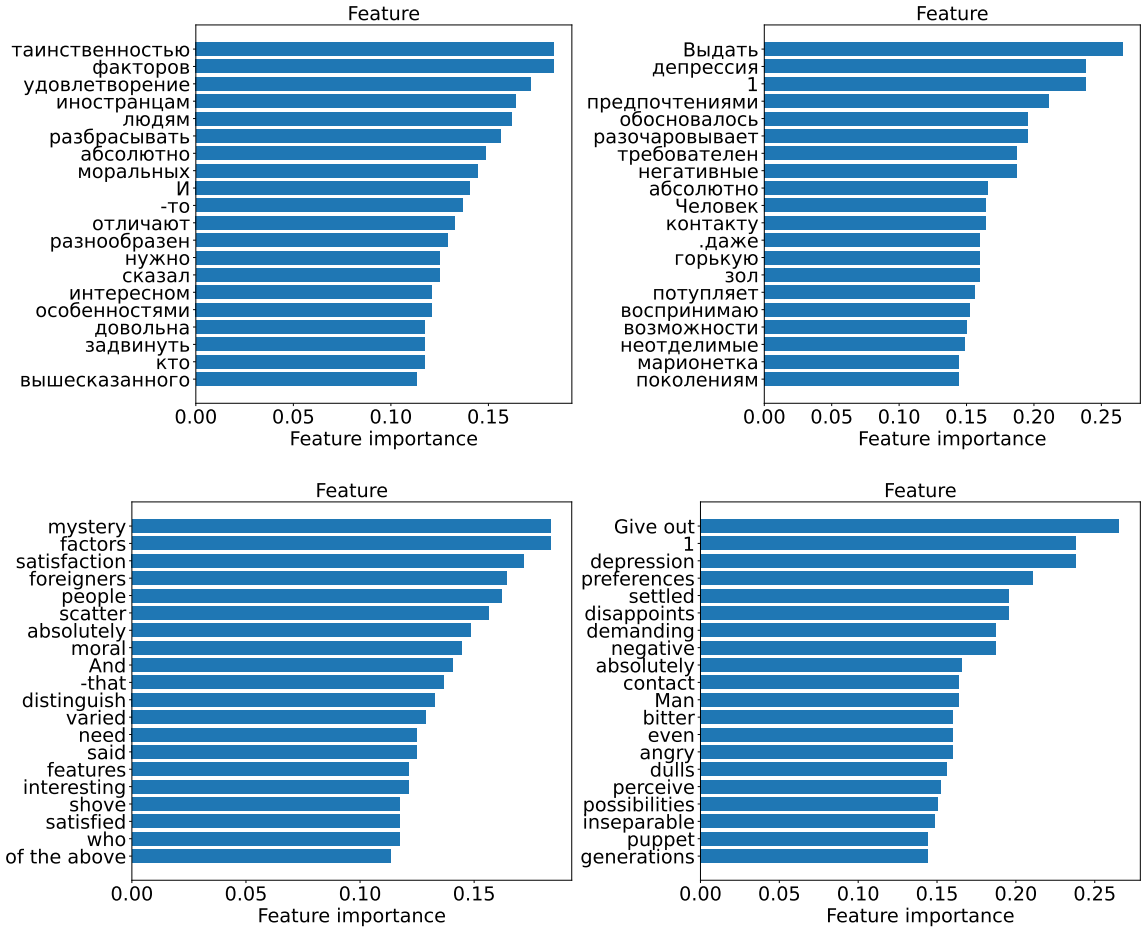

Figure 1: The most important features for the target label generation from the test set of the DE dataset, for the predicted healthy class (left) and pathology class (right). The features are given in Russian (up) and translated into English (down).

Corpus	Model	Mode	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
DE	Gemma2 2B IT	5-shot MMLU	80.00± 0.00	13.33± 0.00	22.86± 0.00	19.59± 0.00	86.36± 0.00	31.93± 0.00	27.39± 0.00
	Gemma2 2B IT	0-shot MMLU	100.00± 0.00	13.33± 0.00	23.53± 0.00	22.00± 0.00	100.00± 0.00	36.07± 0.00	29.80± 0.00
	Gemma2 9B IT	5-shot MMLU	93.11± 0.08	75.11± 0.89	83.15± 0.58	43.16± 0.85	77.27± 0.00	**55.38± 0.71**	**69.26± 0.64**
	Gemma2 9B IT	0-shot MMLU	92.31± 0.00	53.33± 0.00	67.61± 0.00	30.00± 0.00	81.82± 0.00	43.90± 0.00	55.75± 0.00
	Qwen2 7B IT	5-shot MMLU	88.80± 0.17	26.44± 0.44	40.75± 0.55	22.30± 0.10	86.36± 0.00	35.45± 0.13	38.10± 0.34
	Qwen2 7B IT	0-shot MMLU	92.42± 0.16	40.67± 0.89	56.48± 0.89	26.25± 0.29	86.36± 0.00	40.26± 0.34	48.37± 0.62
	SaigaLlama3 8B	5-shot MMLU	85.00± 0.00	18.89± 0.00	30.91± 0.00	20.65± 0.00	86.36± 0.00	33.33± 0.00	32.12± 0.00
	SaigaLlama3 8B	0-shot MMLU	88.37± 0.26	33.78± 0.89	48.87± 0.96	23.20± 0.24	81.82± 0.00	36.15± 0.29	42.51± 0.63
	Vikhr 7B IT 0.4	5-shot MMLU	74.44± 1.11	16.22± 0.89	26.63± 1.28	18.40± 0.16	77.27± 0.00	29.72± 0.21	28.18± 0.74
	Vikhr 7B IT 0.4	0-shot MMLU	0.00± 0.00	0.00± 0.00	0.00± 0.00	19.64± 0.00	100.00± 0.00	32.84± 0.00	16.42± 0.00
	Vikhr 7B IT 5.4	5-shot MMLU	87.06± 0.00	82.22± 0.00	84.57± 0.00	40.74± 0.00	50.00± 0.00	44.90± 0.00	64.73± 0.00
	Vikhr 7B IT 5.4	0-shot MMLU	71.43± 0.00	5.56± 0.00	10.31± 0.00	19.05± 0.00	90.91± 0.00	31.50± 0.00	20.90± 0.00
	VikhrGemma 2B IT	5-shot MMLU	85.45± 0.06	91.33± 0.44	88.29± 0.24	50.67± 1.33	36.36± 0.00	42.33± 0.46	65.31± 0.35
	VikhrGemma 2B IT	0-shot MMLU	80.14± 0.07	98.67± 0.44	88.45± 0.22	0.00± 0.00	0.00± 0.00	0.00± 0.00	44.22± 0.11
DSM	Gemma2 2B IT	5-shot MMLU	71.14± 0.57	20.00± 2.96	31.10± 3.37	42.26± 0.30	87.78± 2.22	57.04± 0.22	44.07± 1.58
	Gemma2 2B IT	0-shot MMLU	60.00± 0.00	44.44± 0.00	51.06± 0.00	40.00± 0.00	55.56± 0.00	46.51± 0.00	48.79± 0.00
	Gemma2 9B IT	5-shot MMLU	64.91± 0.40	82.22± 1.48	72.54± 0.82	55.64± 2.18	33.33± 0.00	41.67± 0.59	57.11± 0.71
	Gemma2 9B IT	0-shot MMLU	64.00± 0.00	59.26± 0.00	61.54± 0.00	45.00± 0.00	50.00± 0.00	47.37± 0.00	54.45± 0.00
	Qwen2 7B IT	5-shot MMLU	66.67± 0.00	44.44± 0.00	53.33± 0.00	44.44± 0.00	66.67± 0.00	53.33± 0.00	53.33± 0.00
	Qwen2 7B IT	0-shot MMLU	62.96± 0.00	62.96± 0.00	62.96± 0.00	44.44± 0.00	44.44± 0.00	44.44± 0.00	53.70± 0.00
	SaigaLlama3 8B	5-shot MMLU	100.00± 0.00	3.70± 0.00	7.14± 0.00	40.91± 0.00	100.00± 0.00	**58.06± 0.00**	32.60± 0.00
	SaigaLlama3 8B	0-shot MMLU	63.64± 0.00	51.85± 0.00	57.14± 0.00	43.48± 0.00	55.56± 0.00	48.78± 0.00	52.96± 0.00
	Vikhr 7B IT 0.4	5-shot MMLU	60.00± 0.00	100.00± 0.00	75.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	37.50± 0.00
	Vikhr 7B IT 0.4	0-shot MMLU	0.00± 0.00	0.00± 0.00	0.00± 0.00	35.71± 0.00	83.33± 0.00	50.00± 0.00	25.00± 0.00
	Vikhr 7B IT 5.4	5-shot MMLU	68.75± 0.00	81.48± 0.00	74.58± 0.00	61.54± 0.00	44.44± 0.00	51.61± 0.00	**63.09± 0.00**
	Vikhr 7B IT 5.4	0-shot MMLU	60.87± 0.00	51.85± 0.00	56.00± 0.00	40.91± 0.00	50.00± 0.00	45.00± 0.00	50.50± 0.00
	VikhrGemma 2B IT	5-shot MMLU	62.34± 0.88	95.56± 1.48	75.44± 0.16	66.67± 0.00	13.33± 4.44	21.90± 5.71	48.67± 2.94
	VikhrGemma 2B IT	0-shot MMLU	60.00± 0.00	100.00± 0.00	75.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	37.50± 0.00
AL	Gemma2 2B IT	5-shot MMLU	50.00± 0.00	4.55± 0.00	8.33± 0.00	46.15± 0.00	94.74± 0.00	62.07± 0.00	35.20± 0.00
	Gemma2 2B IT	0-shot MMLU	0.00± 0.00	0.00± 0.00	0.00± 0.00	46.34± 0.00	100.00± 0.00	**63.33± 0.00**	31.67± 0.00
	Gemma2 9B IT	5-shot MMLU	66.67± 0.00	45.45± 0.00	54.05± 0.00	53.85± 0.00	73.68± 0.00	62.22± 0.00	58.14± 0.00
	Gemma2 9B IT	0-shot MMLU	64.71± 0.00	50.00± 0.00	56.41± 0.00	54.17± 0.00	68.42± 0.00	60.47± 0.00	58.44± 0.00
	Qwen2 7B IT	5-shot MMLU	60.00± 0.00	27.27± 0.00	37.50± 0.00	48.39± 0.00	78.95± 0.00	60.00± 0.00	48.75± 0.00
	Qwen2 7B IT	0-shot MMLU	61.03± 0.52	85.45± 1.82	71.20± 0.99	68.73± 2.55	36.84± 0.00	47.95± 0.64	**59.58± 0.82**
	SaigaLlama3 8B	5-shot MMLU	0.00± 0.00	0.00± 0.00	0.00± 0.00	46.34± 0.00	100.00± 0.00	**63.33± 0.00**	31.67± 0.00
	SaigaLlama3 8B	0-shot MMLU	63.64± 0.00	31.82± 0.00	42.42± 0.00	50.00± 0.00	78.95± 0.00	61.22± 0.00	51.82± 0.00
	Vikhr 7B IT 0.4	5-shot MMLU	55.88± 0.00	86.36± 0.00	67.86± 0.00	57.14± 0.00	21.05± 0.00	30.77± 0.00	49.31± 0.00
	Vikhr 7B IT 0.4	0-shot MMLU	0.00± 0.00	0.00± 0.00	0.00± 0.00	45.00± 0.00	94.74± 0.00	61.02± 0.00	30.51± 0.00
	Vikhr 7B IT 5.4	5-shot MMLU	60.00± 0.00	81.82± 0.00	69.23± 0.00	63.64± 0.00	36.84± 0.00	46.67± 0.00	57.95± 0.00
	Vikhr 7B IT 5.4	0-shot MMLU	33.33± 0.00	4.55± 0.00	8.00± 0.00	44.74± 0.00	89.47± 0.00	59.65± 0.00	33.82± 0.00
	VikhrGemma 2B IT	5-shot MMLU	51.92± 0.60	73.64± 1.82	60.89± 1.03	40.89± 1.78	21.05± 0.00	27.78± 0.39	44.34± 0.71
	VikhrGemma 2B IT	0-shot MMLU	53.66± 0.00	100.00± 0.00	69.84± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	34.92± 0.00
AD	Gemma2 2B IT	5-shot MMLU	0.00± 0.00	0.00± 0.00	0.00± 0.00	47.37± 0.00	100.00± 0.00	64.29± 0.00	32.14± 0.00
	Gemma2 2B IT	0-shot MMLU	50.00± 0.00	10.00± 0.00	16.67± 0.00	47.06± 0.00	88.89± 0.00	61.54± 0.00	39.10± 0.00
	Gemma2 9B IT	5-shot MMLU	55.71± 2.86	19.00± 2.00	28.32± 2.62	48.08± 0.60	83.33± 0.00	60.98± 0.49	44.65± 1.56
	Gemma2 9B IT	0-shot MMLU	77.78± 0.00	35.00± 0.00	48.28± 0.00	55.17± 0.00	88.89± 0.00	68.09± 0.00	58.18± 0.00
	Qwen2 7B IT	5-shot MMLU	100.00± 0.00	5.00± 0.00	9.52± 0.00	48.65± 0.00	100.00± 0.00	65.45± 0.00	37.49± 0.00
	Qwen2 7B IT	0-shot MMLU	70.74± 1.47	70.00± 0.00	70.36± 0.72	67.02± 0.70	67.78± 2.22	67.39± 1.44	**68.87± 1.08**
	SaigaLlama3 8B	5-shot MMLU	0.00± 0.00	0.00± 0.00	0.00± 0.00	47.37± 0.00	100.00± 0.00	64.29± 0.00	32.14± 0.00
	SaigaLlama3 8B	0-shot MMLU	80.00± 0.00	20.00± 0.00	32.00± 0.00	51.52± 0.00	94.44± 0.00	66.67± 0.00	49.33± 0.00
	Vikhr 7B IT 0.4	5-shot MMLU	55.62± 1.25	89.00± 2.00	68.46± 1.54	63.33± 6.67	21.11± 2.22	31.67± 3.33	50.06± 2.44
	Vikhr 7B IT 0.4	0-shot MMLU	13.33± 26.67	2.00± 4.00	3.48± 6.96	46.47± 1.05	94.44± 0.00	62.28± 0.93	32.88± 3.94
	Vikhr 7B IT 5.4	5-shot MMLU	66.67± 0.00	10.00± 0.00	17.39± 0.00	48.57± 0.00	94.44± 0.00	64.15± 0.00	40.77± 0.00
	Vikhr 7B IT 5.4	0-shot MMLU	83.33± 0.00	25.00± 0.00	38.46± 0.00	53.12± 0.00	94.44± 0.00	**68.00± 0.00**	53.23± 0.00
	VikhrGemma 2B IT	5-shot MMLU	38.67± 2.67	10.00± 0.00	15.88± 0.25	45.11± 0.68	82.22± 2.22	58.26± 1.13	37.07± 0.69
	VikhrGemma 2B IT	0-shot MMLU	52.63± 0.00	100.00± 0.00	68.97± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	34.48± 0.00
AC	Gemma2 2B IT	5-shot MMLU	74.55± 0.91	19.56± 0.89	30.98± 1.20	49.16± 0.27	92.11± 0.00	64.10± 0.23	47.54± 0.72
	Gemma2 2B IT	0-shot MMLU	72.73± 0.00	17.78± 0.00	28.57± 0.00	48.61± 0.00	92.11± 0.00	63.64± 0.00	46.10± 0.00
	Gemma2 9B IT	5-shot MMLU	60.46± 1.66	51.56± 18.67	53.91± 7.83	52.31± 4.62	58.95± 18.95	52.50± 10.56	53.20± 1.37
	Gemma2 9B IT	0-shot MMLU	58.46± 1.54	33.78± 0.89	42.82± 1.13	47.72± 0.70	71.58± 1.05	57.26± 0.84	50.04± 0.98
	Qwen2 7B IT	5-shot MMLU	71.52± 2.42	34.67± 1.78	46.69± 2.13	51.97± 0.98	83.68± 1.05	64.12± 1.06	55.41± 1.60
	Qwen2 7B IT	0-shot MMLU	59.89± 0.22	52.44± 1.78	55.91± 1.12	50.93± 0.47	58.42± 1.05	54.41± 0.18	55.16± 0.47
	SaigaLlama3 8B	5-shot MMLU	100.00± 0.00	13.33± 0.00	23.53± 0.00	49.35± 0.00	100.00± 0.00	**66.09± 0.00**	44.81± 0.00
	SaigaLlama3 8B	0-shot MMLU	67.41± 1.48	40.44± 0.89	50.56± 1.11	52.14± 0.71	76.84± 1.05	62.13± 0.85	56.34± 0.98
	Vikhr 7B IT 0.4	5-shot MMLU	54.22± 0.00	100.00± 0.00	70.31± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	35.16± 0.00
	Vikhr 7B IT 0.4	0-shot MMLU	69.15± 2.43	41.78± 0.89	52.09± 1.39	53.04± 1.06	77.89± 2.11	63.11± 1.45	**57.60± 1.42**
	Vikhr 7B IT 5.4	5-shot MMLU	90.67± 18.67	8.89± 4.44	15.33± 5.67	47.12± 0.76	96.32± 7.37	63.22± 2.37	39.28± 1.65
	Vikhr 7B IT 5.4	0-shot MMLU	57.50± 0.00	51.11± 0.00	54.12± 0.00	48.84± 0.00	55.26± 0.00	51.85± 0.00	52.98± 0.00
	VikhrGemma 2B IT	5-shot MMLU	55.29± 0.82	97.78± 0.00	70.63± 0.67	60.00± 30.00	6.32± 3.16	11.43± 5.71	41.03± 3.19
	VikhrGemma 2B IT	0-shot MMLU	53.66± 0.00	97.78± 0.00	69.29± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	34.65± 0.00

Table 12: The results for LLMs MMLU-style evaluation on the five main datasets.

Corpus	Model	Mode	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro	Matched percentage
DE	Gemma2 2B IT	5-shot	82.38± 1.12	10.44± 0.89	18.53± 1.41	20.08± 0.16	90.91± 0.00	32.90± 0.22	17.14± 0.54	99.11± 0.00
	Gemma2 2B IT	0-shot	95.83± 0.00	25.56± 0.00	40.35± 0.00	23.86± 0.00	95.45± 0.00	38.18± 0.00	39.27± 0.00	100.00± 0.00
	Gemma2 9B IT	5-shot	92.80± 0.11	57.33± 0.89	70.88± 0.72	31.92± 0.44	81.82± 0.00	45.92± 0.46	58.40± 0.59	100.00± 0.00
	Gemma2 9B IT	0-shot	93.33± 0.00	62.22± 0.00	74.67± 0.00	34.62± 0.00	81.82± 0.00	**48.65± 0.00**	61.66± 0.00	100.00± 0.00
	Qwen2 7B IT	5-shot	88.80± 0.17	26.44± 0.44	40.75± 0.55	22.30± 0.10	86.36± 0.00	35.45± 0.13	38.10± 0.34	100.00± 0.00
	Qwen2 7B IT	0-shot	92.33± 1.05	42.89± 0.89	58.57± 1.04	26.78± 0.72	85.45± 1.82	40.79± 1.04	49.68± 1.04	100.00± 0.00
	SaigaLlama3 8B	5-shot	85.00± 0.00	18.89± 0.00	30.91± 0.00	20.65± 0.00	86.36± 0.00	33.33± 0.00	32.12± 0.00	100.00± 0.00
	SaigaLlama3 8B	0-shot	100.00± 0.00	8.00± 0.44	14.81± 0.76	20.99± 0.08	100.00± 0.00	34.70± 0.11	24.76± 0.43	100.00± 0.00
	Vikhr 7B IT 0.4	5-shot	73.07± 1.24	15.11± 0.89	25.04± 1.30	18.20± 0.15	77.27± 0.00	29.46± 0.20	27.25± 0.75	100.00± 0.00
	Vikhr 7B IT 0.4	0-shot	0.00± 0.00	0.00± 0.00	0.00± 0.00	19.64± 0.00	100.00± 0.00	32.84± 0.00	16.42± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	5-shot	87.06± 0.00	82.22± 0.00	84.57± 0.00	40.74± 0.00	50.00± 0.00	44.90± 0.00	**64.73± 0.00**	100.00± 0.00
	Vikhr 7B IT 5.4	0-shot	71.43± 0.00	5.56± 0.00	10.31± 0.00	19.05± 0.00	90.91± 0.00	31.50± 0.00	20.90± 0.00	100.00± 0.00
	VikhrGemma 2B IT	5-shot	88.54± 0.07	68.67± 0.44	77.35± 0.31	33.18± 0.31	63.64± 0.00	43.62± 0.27	60.48± 0.29	100.00± 0.00
	VikhrGemma 2B IT	0-shot	85.24± 0.95	6.44± 0.44	11.98± 0.78	20.59± 0.40	94.55± 1.82	33.82± 0.65	15.27± 0.48	96.25± 0.36
DSM	Gemma2 2B IT	5-shot	54.00± 8.00	11.11± 7.41	17.89± 9.97	39.45± 0.86	86.67± 4.44	54.16± 0.16	32.97± 1.20	99.56± 0.89
	Gemma2 2B IT	0-shot	64.71± 0.00	40.74± 0.00	50.00± 0.00	42.86± 0.00	66.67± 0.00	52.17± 0.00	51.09± 0.00	100.00± 0.00
	Gemma2 9B IT	5-shot	64.24± 1.21	29.63± 7.41	40.14± 6.60	41.83± 1.31	75.56± 4.44	53.74± 0.21	46.94± 3.20	100.00± 0.00
	Gemma2 9B IT	0-shot	62.64± 0.44	80.74± 1.48	70.54± 0.85	49.09± 1.82	27.78± 0.00	35.47± 0.49	53.01± 0.67	100.00± 0.00
	Qwen2 7B IT	5-shot	70.59± 0.00	44.44± 0.00	54.55± 0.00	46.43± 0.00	72.22± 0.00	56.52± 0.00	55.53± 0.00	100.00± 0.00
	Qwen2 7B IT	0-shot	65.52± 0.00	70.37± 0.00	67.86± 0.00	50.00± 0.00	44.44± 0.00	47.06± 0.00	57.46± 0.00	100.00± 0.00
	SaigaLlama3 8B	5-shot	100.00± 0.00	3.70± 0.00	7.14± 0.00	40.91± 0.00	100.00± 0.00	**58.06± 0.00**	32.60± 0.00	100.00± 0.00
	SaigaLlama3 8B	0-shot	55.30± 1.52	22.96± 1.48	32.44± 1.73	38.47± 0.46	72.22± 0.00	50.20± 0.39	41.32± 1.06	100.00± 0.00
	Vikhr 7B IT 0.4	5-shot	69.19± 0.89	81.48± 0.00	74.83± 0.51	62.09± 1.10	45.56± 2.22	52.54± 1.85	**63.69± 1.18**	100.00± 0.00
	Vikhr 7B IT 0.4	0-shot	0.00± 0.00	0.00± 0.00	0.00± 0.00	35.71± 0.00	83.33± 0.00	50.00± 0.00	25.00± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	5-shot	68.75± 0.00	81.48± 0.00	74.58± 0.00	61.54± 0.00	44.44± 0.00	51.61± 0.00	63.09± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	0-shot	59.09± 0.00	48.15± 0.00	53.06± 0.00	39.13± 0.00	50.00± 0.00	43.90± 0.00	48.48± 0.00	100.00± 0.00
	VikhrGemma 2B IT	5-shot	61.34± 5.04	37.04± 0.00	46.12± 1.33	40.46± 2.35	64.44± 6.67	49.69± 3.73	47.90± 2.53	100.00± 0.00
	VikhrGemma 2B IT	0-shot	57.14± 0.00	14.81± 0.00	23.53± 0.00	41.94± 0.00	72.22± 0.00	53.06± 0.00	25.53± 0.00	84.44± 0.00
AL	Gemma2 2B IT	5-shot	57.14± 0.00	18.18± 0.00	27.59± 0.00	47.06± 0.00	84.21± 0.00	60.38± 0.00	43.98± 0.00	100.00± 0.00
	Gemma2 2B IT	0-shot	63.64± 0.00	31.82± 0.00	42.42± 0.00	50.00± 0.00	78.95± 0.00	61.22± 0.00	51.82± 0.00	100.00± 0.00
	Gemma2 9B IT	5-shot	75.00± 0.00	27.27± 0.00	40.00± 0.00	51.52± 0.00	89.47± 0.00	**65.38± 0.00**	52.69± 0.00	100.00± 0.00
	Gemma2 9B IT	0-shot	66.67± 0.00	63.64± 0.00	65.12± 0.00	60.00± 0.00	63.16± 0.00	61.54± 0.00	**63.33± 0.00**	100.00± 0.00
	Qwen2 7B IT	5-shot	61.33± 2.67	27.27± 0.00	37.74± 0.48	48.71± 0.65	80.00± 2.11	60.55± 1.10	49.15± 0.79	100.00± 0.00
	Qwen2 7B IT	0-shot	60.00± 0.00	95.45± 0.00	73.68± 0.00	83.33± 0.00	26.32± 0.00	40.00± 0.00	56.84± 0.00	100.00± 0.00
	SaigaLlama3 8B	5-shot	100.00± 0.00	4.55± 0.00	8.70± 0.00	47.50± 0.00	100.00± 0.00	64.41± 0.00	36.55± 0.00	100.00± 0.00
	SaigaLlama3 8B	0-shot	36.67± 6.67	5.45± 1.82	9.48± 2.95	44.98± 0.48	89.47± 0.00	59.86± 0.43	34.67± 1.69	100.00± 0.00
	Vikhr 7B IT 0.4	5-shot	50.00± 0.00	45.45± 0.00	47.62± 0.00	42.86± 0.00	47.37± 0.00	45.00± 0.00	46.31± 0.00	100.00± 0.00
	Vikhr 7B IT 0.4	0-shot	0.00± 0.00	0.00± 0.00	0.00± 0.00	45.00± 0.00	94.74± 0.00	61.02± 0.00	30.51± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	5-shot	60.00± 0.00	81.82± 0.00	69.23± 0.00	63.64± 0.00	36.84± 0.00	46.67± 0.00	57.95± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	0-shot	33.33± 0.00	4.55± 0.00	8.00± 0.00	44.74± 0.00	89.47± 0.00	59.65± 0.00	33.82± 0.00	100.00± 0.00
	VikhrGemma 2B IT	5-shot	66.67± 0.00	27.27± 0.00	38.71± 0.00	50.00± 0.00	84.21± 0.00	62.75± 0.00	50.73± 0.00	100.00± 0.00
	VikhrGemma 2B IT	0-shot	0.00± 0.00	0.00± 0.00	0.00± 0.00	43.87± 0.56	90.53± 2.11	59.10± 0.96	29.55± 0.48	100.00± 0.00
AD	Gemma2 2B IT	5-shot	0.00± 0.00	0.00± 0.00	0.00± 0.00	47.37± 0.00	100.00± 0.00	64.29± 0.00	32.14± 0.00	100.00± 0.00
	Gemma2 2B IT	0-shot	64.29± 0.00	45.00± 0.00	52.94± 0.00	54.17± 0.00	72.22± 0.00	61.90± 0.00	57.42± 0.00	100.00± 0.00
	Gemma2 9B IT	5-shot	100.00± 0.00	5.00± 0.00	9.52± 0.00	48.65± 0.00	100.00± 0.00	65.45± 0.00	37.49± 0.00	100.00± 0.00
	Gemma2 9B IT	0-shot	60.87± 0.00	70.00± 0.00	65.12± 0.00	60.00± 0.00	50.00± 0.00	54.55± 0.00	59.83± 0.00	100.00± 0.00
	Qwen2 7B IT	5-shot	100.00± 0.00	5.00± 0.00	9.52± 0.00	48.65± 0.00	100.00± 0.00	65.45± 0.00	37.49± 0.00	100.00± 0.00
	Qwen2 7B IT	0-shot	67.25± 1.16	80.00± 0.00	73.07± 0.68	71.81± 0.76	56.67± 2.22	63.33± 1.67	**68.20± 1.17**	100.00± 0.00
	SaigaLlama3 8B	5-shot	100.00± 0.00	5.00± 0.00	9.52± 0.00	48.65± 0.00	100.00± 0.00	65.45± 0.00	37.49± 0.00	100.00± 0.00
	SaigaLlama3 8B	0-shot	54.55± 0.00	30.00± 0.00	38.71± 0.00	48.15± 0.00	72.22± 0.00	57.78± 0.00	48.24± 0.00	100.00± 0.00
	Vikhr 7B IT 0.4	5-shot	50.91± 1.82	30.00± 0.00	37.74± 0.48	46.55± 0.80	67.78± 2.22	55.19± 1.29	46.47± 0.89	100.00± 0.00
	Vikhr 7B IT 0.4	0-shot	13.33± 26.67	2.00± 4.00	3.48± 6.96	46.47± 1.05	94.44± 0.00	62.28± 0.93	32.88± 3.94	100.00± 0.00
	Vikhr 7B IT 5.4	5-shot	66.67± 0.00	10.00± 0.00	17.39± 0.00	48.57± 0.00	94.44± 0.00	64.15± 0.00	40.77± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	0-shot	83.33± 0.00	25.00± 0.00	38.46± 0.00	53.12± 0.00	94.44± 0.00	**68.00± 0.00**	53.23± 0.00	100.00± 0.00
	VikhrGemma 2B IT	5-shot	100.00± 0.00	5.00± 0.00	9.52± 0.00	48.65± 0.00	100.00± 0.00	65.45± 0.00	37.49± 0.00	100.00± 0.00
	VikhrGemma 2B IT	0-shot	66.67± 0.00	20.00± 0.00	30.77± 0.00	50.00± 0.00	88.89± 0.00	64.00± 0.00	47.38± 0.00	100.00± 0.00
AC	Gemma2 2B IT	5-shot	72.00± 2.67	24.00± 0.89	36.00± 1.33	50.45± 0.60	88.95± 1.05	64.38± 0.76	33.46± 0.70	98.80± 0.00
	Gemma2 2B IT	0-shot	97.50± 5.00	15.56± 0.00	26.82± 0.20	49.87± 0.27	99.47± 1.05	66.43± 0.47	46.63± 0.34	100.00± 0.00
	Gemma2 9B IT	5-shot	68.00± 16.00	13.78± 0.89	22.84± 2.04	47.26± 1.37	91.58± 4.21	62.34± 2.16	42.59± 2.10	100.00± 0.00
	Gemma2 9B IT	0-shot	54.55± 0.00	26.67± 0.00	35.82± 0.00	47.46± 0.00	73.68± 0.00	57.73± 0.00	31.18± 0.00	97.59± 0.00
	Qwen2 7B IT	5-shot	84.89± 3.56	26.22± 5.33	39.85± 6.96	52.08± 1.72	94.74± 0.00	**67.20± 1.46**	53.52± 4.21	100.00± 0.00
	Qwen2 7B IT	0-shot	55.93± 0.24	52.44± 1.78	54.11± 0.87	47.55± 0.09	51.05± 2.11	49.22± 1.00	51.67± 0.07	100.00± 0.00
	SaigaLlama3 8B	5-shot	100.00± 0.00	13.33± 0.00	23.53± 0.00	49.35± 0.00	100.00± 0.00	66.09± 0.00	44.81± 0.00	100.00± 0.00
	SaigaLlama3 8B	0-shot	66.05± 0.85	46.67± 0.00	54.69± 0.29	53.12± 0.36	71.58± 1.05	60.98± 0.62	**57.84± 0.45**	100.00± 0.00
	Vikhr 7B IT 0.4	5-shot	54.20± 0.30	88.89± 0.00	67.34± 0.23	45.56± 2.22	11.05± 1.05	17.78± 1.52	42.56± 0.88	100.00± 0.00
	Vikhr 7B IT 0.4	0-shot	67.48± 1.05	37.78± 0.00	48.43± 0.27	51.55± 0.34	78.42± 1.05	62.21± 0.58	55.32± 0.43	100.00± 0.00
	Vikhr 7B IT 5.4	5-shot	90.67± 18.67	8.89± 4.44	15.33± 5.67	47.12± 0.76	96.32± 7.37	63.22± 2.37	39.28± 1.65	100.00± 0.00
	Vikhr 7B IT 5.4	0-shot	57.50± 0.00	51.11± 0.00	54.12± 0.00	48.84± 0.00	55.26± 0.00	51.85± 0.00	52.98± 0.00	100.00± 0.00
	VikhrGemma 2B IT	5-shot	52.75± 1.02	80.00± 4.44	63.56± 2.12	40.00± 5.00	15.26± 1.05	21.94± 0.56	42.75± 0.78	100.00± 0.00
	VikhrGemma 2B IT	0-shot	64.26± 0.88	24.00± 0.89	34.94± 1.08	50.00± 0.00	73.68± 0.00	59.57± 0.00	31.51± 0.36	87.71± 0.48

Table 13: The results for LLMs evaluation on the five main datasets.

Corpus	Model	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
D-all	Gemma2 2B IT	92.33± 0.94	97.14± 1.04	94.67± 0.72	76.44± 6.36	52.98± 6.33	**62.33± 5.60**	**78.50± 3.13**
	Gemma2 9B IT	91.58± 1.24	97.65± 0.82	94.51± 0.68	77.86± 6.23	47.62± 8.42	58.69± 6.91	76.60± 3.77
	Qwen2 7B IT	91.09± 0.64	97.03± 0.96	93.96± 0.17	72.97± 5.26	44.64± 4.94	54.98± 2.77	74.47± 1.36
	SaigaLlama3 8B	90.88± 1.00	96.73± 1.49	93.71± 1.01	70.16± 10.04	43.45± 6.65	53.41± 7.35	73.56± 4.13
	Vikhr 7B IT 0.4	90.46± 1.23	96.83± 1.20	93.53± 1.01	68.86± 11.28	40.48± 8.16	50.73± 8.61	72.13± 4.79
	Vikhr 7B IT 5.4	90.18± 1.16	92.02± 1.28	91.09± 1.16	47.32± 7.90	41.67± 7.04	44.28± 7.35	67.69± 4.24
	VikhrGemma 2B IT	91.83± 0.80	97.44± 1.68	94.54± 0.66	79.25± 11.48	49.40± 5.98	60.04± 3.80	77.29± 2.09
	RuRoBERTa	92.48± 0.63	95.50± 1.16	93.96± 0.42	68.19± 5.05	54.76± 4.45	60.48± 2.31	77.22± 1.27
	RuBioRoBERTa	90.74± 2.52	96.93± 1.66	93.69± 1.04	58.50± 26.99	41.67± 19.20	48.61± 22.33	71.15± 11.60
	RuBERT	92.26± 0.52	96.22± 1.56	94.19± 0.68	72.12± 10.10	52.98± 3.81	60.60± 2.87	77.39± 1.72
	BERT	90.16± 3.31	96.63± 1.69	93.22± 1.43	50.52± 28.15	37.50± 24.38	42.50± 26.53	67.86± 13.90
	Linguistic features	88.10± 0.60	88.10± 3.40	88.00± 1.60	51.60± 5.80	50.70± 4.10	50.80± 2.60	69.40± 1.90
	TF-IDF	89.50± 1.00	84.20± 1.00	86.80± 0.60	47.50± 1.80	59.10± 4.20	52.60± 2.50	69.70± 1.50
A-all	Gemma2 2B IT	57.61± 3.21	57.09± 4.28	57.27± 3.15	50.61± 3.62	51.11± 6.00	50.77± 4.48	54.02± 3.39
	Gemma2 9B IT	55.44± 1.46	56.32± 4.04	55.83± 2.54	48.47± 1.77	47.56± 3.14	47.95± 1.96	51.89± 1.63
	Qwen2 7B IT	56.43± 1.57	56.13± 3.95	56.23± 2.59	49.51± 2.04	49.78± 2.85	49.59± 1.86	52.91± 1.79
	SaigaLlama3 8B	58.56± 1.32	56.32± 4.83	57.33± 2.79	51.59± 2.03	53.78± 3.74	52.56± 1.83	54.95± 1.57
	Vikhr 7B IT 0.4	57.08± 3.28	56.13± 2.43	56.58± 2.68	49.92± 3.40	50.89± 4.95	50.38± 4.09	53.48± 3.32
	Vikhr 7B IT 5.4	56.89± 2.40	55.36± 7.22	55.95± 4.71	50.10± 3.07	51.56± 4.79	50.65± 2.68	53.30± 2.82
	VikhrGemma 2B IT	54.79± 2.65	53.26± 2.94	53.97± 2.34	47.34± 2.63	48.89± 4.91	48.06± 3.59	51.01± 2.62
	RuRoBERTa	61.53± 3.69	60.15± 13.80	59.68± 4.47	54.39± 0.97	54.44± 18.08	52.27± 11.87	**55.97± 3.88**
	RuBioRoBERTa	55.16± 1.73	73.18± 19.96	61.66± 6.68	32.94± 23.36	30.22± 22.11	31.31± 22.37	46.48± 8.32
	RuBERT	57.18± 1.93	51.53± 5.61	54.09± 3.67	49.69± 1.98	55.33± 4.34	52.25± 2.28	53.17± 2.10
	BERT	59.44± 1.74	46.93± 3.66	52.33± 2.10	50.43± 0.98	62.67± 4.81	**55.82± 2.21**	54.08± 1.08
	Linguistic features	53.40± 2.90	39.70± 19.30	42.00± 17.90	46.10± 2.40	60.20± 20.40	51.00± 7.80	46.50± 6.20
	TF-IDF	51.80± 0.70	40.00± 15.60	43.30± 10.80	44.30± 2.10	56.40± 17.80	48.60± 8.50	46.00± 2.40

Table 14: Results of classification on D-all and A-all datasets for encoder models, AutoML models and LLMs with LoRA.

Corpus	Model	Mode	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
D-all	Gemma2 2B IT	5-shot MMLU	100.00± 0.00	7.36± 0.00	13.71± 0.00	15.64± 0.00	100.00± 0.00	27.05± 0.00	20.38± 0.00
	Gemma2 2B IT	0-shot MMLU	100.00± 0.00	30.55± 0.25	46.80± 0.29	19.83± 0.06	100.00± 0.00	33.10± 0.08	39.95± 0.18
	Gemma2 9B IT	5-shot MMLU	88.83± 0.25	96.56± 0.49	92.53± 0.36	59.56± 4.84	29.29± 1.43	39.26± 2.32	**65.90± 1.34**
	Gemma2 9B IT	0-shot MMLU	94.87± 0.00	68.10± 0.00	79.29± 0.00	29.73± 0.00	78.57± 0.00	**43.14± 0.00**	61.21± 0.00
	Qwen2 7B IT	5-shot MMLU	94.74± 0.00	22.09± 0.00	35.82± 0.00	16.99± 0.00	92.86± 0.00	28.73± 0.00	32.28± 0.00
	Qwen2 7B IT	0-shot MMLU	95.66± 0.08	54.11± 0.98	69.12± 0.83	24.30± 0.38	85.71± 0.00	37.86± 0.47	53.49± 0.65
	SaigaLlama3 8B	5-shot MMLU	100.00± 0.00	1.10± 0.25	2.18± 0.48	14.80± 0.03	100.00± 0.00	25.78± 0.05	13.98± 0.26
	SaigaLlama3 8B	0-shot MMLU	92.98± 0.04	40.61± 0.25	56.53± 0.24	19.20± 0.06	82.14± 0.00	31.12± 0.08	43.83± 0.16
	Vikhr 7B IT 0.4	5-shot MMLU	88.97± 0.24	90.06± 0.25	89.51± 0.24	37.69± 1.54	35.00± 1.43	36.30± 1.48	62.90± 0.86
	Vikhr 7B IT 0.4	0-shot MMLU	100.00± 0.00	2.45± 0.00	4.79± 0.00	14.97± 0.00	100.00± 0.00	26.05± 0.00	15.42± 0.00
	Vikhr 7B IT 5.4	5-shot MMLU	96.88± 0.00	57.06± 0.00	71.81± 0.00	26.32± 0.00	89.29± 0.00	40.65± 0.00	56.23± 0.00
	Vikhr 7B IT 5.4	0-shot MMLU	90.94± 0.06	30.80± 0.25	46.01± 0.28	16.94± 0.05	82.14± 0.00	28.08± 0.07	37.05± 0.18
	VikhrGemma 2B IT	5-shot MMLU	84.93± 0.00	38.04± 0.00	52.54± 0.00	14.41± 0.00	60.71± 0.00	23.29± 0.00	37.92± 0.00
	VikhrGemma 2B IT	0-shot MMLU	85.25± 0.03	99.26± 0.25	91.72± 0.12	0.00± 0.00	0.00± 0.00	0.00± 0.00	45.86± 0.06
A-all	Gemma2 2B IT	5-shot MMLU	62.86± 0.71	6.67± 1.84	11.99± 2.92	46.86± 0.22	95.47± 1.07	62.86± 0.04	37.42± 1.44
	Gemma2 2B IT	0-shot MMLU	66.67± 0.00	11.49± 0.00	19.61± 0.00	47.62± 0.00	93.33± 0.00	63.06± 0.00	41.34± 0.00
	Gemma2 9B IT	5-shot MMLU	57.52± 0.19	63.45± 5.06	60.24± 2.04	51.94± 1.03	45.60± 4.80	48.36± 2.58	54.30± 0.27
	Gemma2 9B IT	0-shot MMLU	63.85± 0.77	38.16± 0.46	47.77± 0.58	51.09± 0.36	74.93± 0.53	60.76± 0.43	54.26± 0.50
	Qwen2 7B IT	5-shot MMLU	65.42± 1.30	45.29± 5.52	53.28± 3.06	53.22± 0.77	72.00± 5.33	61.08± 1.61	57.18± 0.73
	Qwen2 7B IT	0-shot MMLU	62.67± 0.06	64.83± 1.38	63.72± 0.65	57.51± 0.47	55.20± 1.07	56.32± 0.32	**60.02± 0.17**
	SaigaLlama3 8B	5-shot MMLU	80.51± 5.64	11.26± 0.46	19.76± 0.88	48.46± 0.41	96.80± 1.07	**64.59± 0.60**	42.18± 0.74
	SaigaLlama3 8B	0-shot MMLU	67.91± 0.93	33.56± 0.46	44.92± 0.62	51.43± 0.34	81.60± 0.53	63.09± 0.41	54.01± 0.51
	Vikhr 7B IT 0.4	5-shot MMLU	56.68± 1.42	80.00± 8.28	66.10± 1.52	54.26± 2.13	28.53± 12.27	35.58± 14.09	50.84± 6.29
	Vikhr 7B IT 0.4	0-shot MMLU	64.91± 1.21	22.07± 0.46	32.93± 0.34	48.79± 0.16	86.13± 1.07	62.29± 0.41	47.61± 0.04
	Vikhr 7B IT 5.4	5-shot MMLU	62.12± 4.24	22.07± 4.14	32.16± 4.54	48.08± 0.10	83.73± 4.80	61.04± 1.30	46.60± 1.62
	Vikhr 7B IT 5.4	0-shot MMLU	59.18± 0.00	33.33± 0.00	42.65± 0.00	48.67± 0.00	73.33± 0.00	58.51± 0.00	50.58± 0.00
	VikhrGemma 2B IT	5-shot MMLU	53.67± 0.50	97.93± 1.84	69.32± 0.06	11.67± 23.33	1.87± 3.73	3.22± 6.44	36.27± 3.19
	VikhrGemma 2B IT	0-shot MMLU	53.42± 0.00	98.85± 0.00	69.35± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	34.68± 0.00

Table 15: The results for LLMs MMLU-style evaluation, D-all and A-all datasets.

Corpus	Model	Mode	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro	Matched percentage
D-all	Gemma2 2B IT	5-shot	94.25± 0.07	20.12± 0.25	33.16± 0.34	16.65± 0.04	92.86± 0.00	28.23± 0.06	30.70± 0.20	100.00± 0.00
	Gemma2 2B IT	0-shot	96.06± 0.56	41.84± 0.25	58.29± 0.34	21.00± 0.33	90.00± 1.43	34.05± 0.54	46.17± 0.44	100.00± 0.00
	Gemma2 9B IT	5-shot	98.17± 0.03	32.88± 0.49	49.26± 0.56	19.80± 0.12	96.43± 0.00	32.85± 0.16	41.06± 0.36	100.00± 0.00
	Gemma2 9B IT	0-shot	93.28± 0.00	76.69± 0.00	84.18± 0.00	33.33± 0.00	67.86± 0.00	**44.71± 0.00**	**64.44± 0.00**	100.00± 0.00
	Qwen2 7B IT	5-shot	94.59± 0.00	21.47± 0.00	35.00± 0.00	16.88± 0.00	92.86± 0.00	28.57± 0.00	31.79± 0.00	100.00± 0.00
	Qwen2 7B IT	0-shot	95.64± 0.48	56.56± 0.98	71.08± 0.91	25.17± 0.73	85.00± 1.43	38.84± 1.02	54.96± 0.96	100.00± 0.00
	SaigaLlama3 8B	5-shot	100.00± 0.00	7.48± 0.98	13.91± 1.72	15.66± 0.14	100.00± 0.00	27.08± 0.21	20.50± 0.97	100.00± 0.00
	SaigaLlama3 8B	0-shot	96.18± 0.06	15.46± 0.25	26.64± 0.37	16.38± 0.04	96.43± 0.00	28.01± 0.06	27.32± 0.21	100.00± 0.00
	Vikhr 7B IT 0.4	5-shot	93.02± 0.08	40.86± 0.49	56.78± 0.49	19.26± 0.13	82.14± 0.00	31.21± 0.17	43.99± 0.33	100.00± 0.00
	Vikhr 7B IT 0.4	0-shot	100.00± 0.00	2.45± 0.00	4.79± 0.00	14.97± 0.00	100.00± 0.00	26.05± 0.00	15.42± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	5-shot	96.88± 0.00	57.06± 0.00	71.81± 0.00	26.32± 0.00	89.29± 0.00	40.65± 0.00	56.23± 0.00	100.00± 0.00
	Vikhr 7B IT 5.4	0-shot	90.94± 0.06	30.80± 0.25	46.01± 0.28	16.94± 0.05	82.14± 0.00	28.08± 0.07	37.05± 0.18	100.00± 0.00
	VikhrGemma 2B IT	5-shot	93.90± 0.08	18.90± 0.25	31.46± 0.35	16.43± 0.04	92.86± 0.00	27.93± 0.06	29.69± 0.20	100.00± 0.00
	VikhrGemma 2B IT	0-shot	96.73± 0.09	18.16± 0.49	30.58± 0.70	17.59± 0.35	95.71± 1.43	29.71± 0.57	20.10± 0.43	95.81± 0.00
A-all	Gemma2 2B IT	5-shot	69.78± 1.11	20.69± 0.00	31.92± 0.11	49.63± 0.00	89.60± 0.53	63.88± 0.14	35.14± 6.50	99.51± 0.25
	Gemma2 2B IT	0-shot	71.44± 0.87	26.44± 0.00	38.59± 0.13	50.69± 0.15	87.73± 0.53	**64.26± 0.27**	51.42± 0.20	100.00± 0.00
	Gemma2 9B IT	5-shot	67.80± 0.57	38.62± 3.68	49.09± 2.66	52.52± 0.70	78.67± 2.67	62.96± 0.40	56.02± 1.13	100.00± 0.00
	Gemma2 9B IT	0-shot	60.61± 0.00	45.98± 0.00	52.29± 0.00	52.13± 0.00	65.33± 0.00	57.99± 0.00	36.76± 0.00	98.77± 0.00
	Qwen2 7B IT	5-shot	67.93± 2.23	34.25± 1.84	45.47± 1.01	51.52± 0.30	81.07± 3.20	62.98± 1.22	54.22± 0.11	100.00± 0.00
	Qwen2 7B IT	0-shot	60.01± 0.41	69.66± 0.92	64.47± 0.17	56.72± 0.10	46.13± 1.60	50.87± 0.99	**57.67± 0.41**	100.00± 0.00
	SaigaLlama3 8B	5-shot	69.34± 0.46	17.70± 1.38	28.18± 1.83	48.79± 0.27	90.93± 0.53	63.50± 0.10	45.84± 0.96	100.00± 0.00
	SaigaLlama3 8B	0-shot	61.30± 0.87	32.41± 0.46	42.41± 0.60	49.31± 0.34	76.27± 0.53	59.90± 0.42	51.15± 0.51	100.00± 0.00
	Vikhr 7B IT 0.4	5-shot	56.94± 1.27	68.05± 5.06	61.86± 1.14	51.63± 1.49	40.00± 8.00	44.69± 6.27	53.28± 2.56	100.00± 0.00
	Vikhr 7B IT 0.4	0-shot	63.04± 0.15	20.00± 0.92	30.36± 1.06	48.22± 0.13	86.40± 0.53	61.89± 0.03	46.12± 0.52	100.00± 0.00
	Vikhr 7B IT 5.4	5-shot	62.12± 4.24	22.07± 4.14	32.16± 4.54	48.08± 0.10	83.73± 4.80	61.04± 1.30	46.60± 1.62	100.00± 0.00
	Vikhr 7B IT 5.4	0-shot	59.18± 0.00	33.33± 0.00	42.65± 0.00	48.67± 0.00	73.33± 0.00	58.51± 0.00	50.58± 0.00	100.00± 0.00
	VikhrGemma 2B IT	5-shot	51.38± 2.77	24.14± 9.20	32.22± 8.23	46.04± 1.46	74.40± 5.87	56.71± 0.87	41.01± 3.23	99.88± 0.25
	VikhrGemma 2B IT	0-shot	60.17± 0.35	17.01± 0.46	26.52± 0.53	48.11± 0.16	81.60± 0.53	60.53± 0.28	29.02± 0.09	93.70± 0.25

Table 16: The results for LLMs evaluation, D-all and A-all datasets.

Error name	Description	Example
1. Tautology	The final part of the explanation repeats the initial part in other words without any proof.	This text **indicates that the author has depression**, as he describes his thoughts and feelings, which are **characteristic of depressive disorders**.
2. Groundless generalization	The patient's experience from the text is defined as a sign of depression, while this experience on its own, without the context, is not specific to depression.	**The desire to return to the lost state of happiness and happy life**, which is also a **sign of depression**.
3. False conclusion	The false inference is derived from the text statement.	The author also mentions that his parents, who he considers to be positive, **were unable to correct his behavior**. *In the original text there is no signs of parents intention to correct authors behaviour.*
4. Confabulation	The explanation contains evidence, which is not mentioned in the context.	The text describes a **deep dissatisfaction with the world, people and their actions** *included in the text*, and also **expresses a desire to be happy** *not in the text* and **enjoy the little things** *not in the text*.
5. Distortion of medical understanding of depression	Misconception about depression.	She also expresses a desire to ... **plan long-term plans, which also indicates the presence of a depressive disorder**. *The long-term planning cannot be considered as a sign of depression.*
6. Incompleteness of selected signs of depression	Of the several significant signs of depression only one or two signs (mostly minor) are highlighted.	Thank you for this test, so that I could repeat all this to myself once again and **think about the rope**. *This significant sign does not mentioned in the explanation.*

Table 17: Detailed description of errors in LLM explanations from the perspective of trained psychologists. The bold font indicates the significant part of the examples, illustrating the error type; the italic font highlights the psychologist's notes for the examples.

Corpus	Model	Mode	Precision healthy	Recall healthy	F1-healthy	Precision pathology	Recall pathology	F1-pathology	F1-macro
DE	0-shot	Vikhr 7B IT 5.4	71.43± 0.00	5.56± 0.00	10.31± 0.00	19.05± 0.00	90.91± 0.00	31.50± 0.00	20.90± 0.00
	5-shot	Vikhr 7B IT 5.4	87.06± 0.00	82.22± 0.00	84.57± 0.00	40.74± 0.00	50.00± 0.00	44.90± 0.00	64.73± 0.00
	5-shot clinically informed	Vikhr 7B IT 5.4	87.64± 0.00	86.67± 0.00	87.15± 0.00	47.83± 0.00	50.00± 0.00	48.89± 0.00	68.02± 0.00

Table 18: Comparison of the results of the best generative model on DE in various settings (mean ± std).

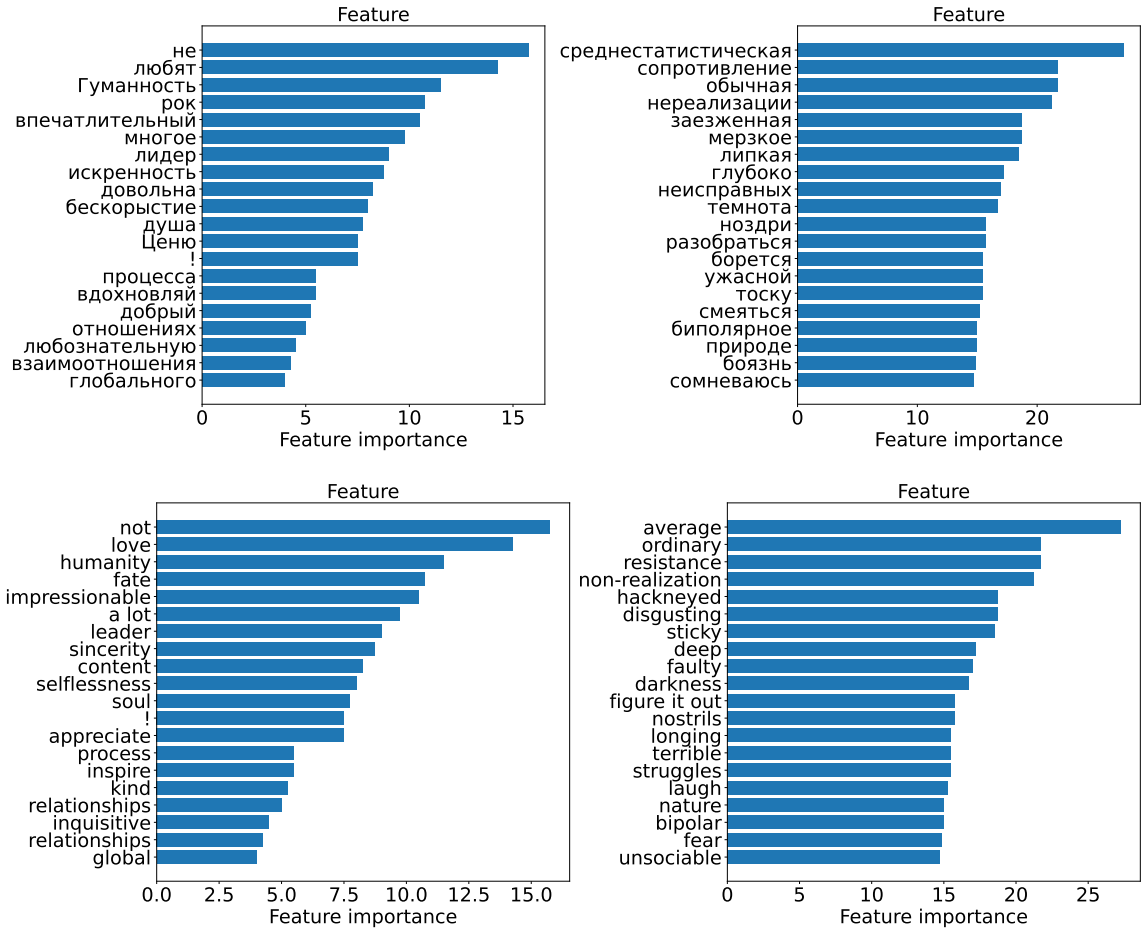

Figure 2: The most important features for the explanation generation from the test set of the DE dataset, for the predicted healthy class (left) and pathology class (right). The features are given in Russian (up) and translated into English (down).

Mode	Model	BDI Score 0-3			BDI Score 3-6			BDI Score 7-10		
		Precision healthy	Recall healthy	F1-healthy	Precision healthy	Recall healthy	F1-healthy	Precision healthy	Recall healthy	F1-healthy
5-shot	SaigaLlama3 8B	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	100.00± 0.00	6.25± 0.00	11.76± 0.00
5-shot	Vikhr 7B IT 0.4	100.00± 0.00	83.33± 0.00	90.91± 0.00	100.00± 0.00	80.00± 0.00	88.89± 0.00	100.00± 0.00	81.25± 0.00	89.66± 0.00
5-shot MMLU	SaigaLlama3 8B	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	0.00± 0.00	100.00± 0.00	6.25± 0.00	11.76± 0.00
5-shot MMLU	Vikhr 7B IT 5.4	100.00± 0.00	83.33± 0.00	90.91± 0.00	100.00± 0.00	80.00± 0.00	88.89± 0.00	100.00± 0.00	81.25± 0.00	89.66± 0.00
LoRA	Qwen2 7B IT	100.00± 0.00	80.56± 11.45	88.79± 7.00	100.00± 0.00	83.33± 21.34	89.15± 15.15	100.00± 0.00	65.62± 10.05	78.80± 7.34
SFT	RuBioRoBERTa	100.00± 0.00	86.11± 11.45	92.12± 6.78	100.00± 0.00	50.00± 10.00	66.07± 8.93	100.00± 0.00	58.33± 4.66	73.57± 3.79

Table 19: Results of best models for DSM dataset for various sub-splits for healthy group (mean ± std).

Mode	Model	BDI Score 30-34			BDI Score 34-63		
		Precision pathology	Recall pathology	F1-pathology	Precision pathology	Recall pathology	F1-pathology
5-shot	SaigaLlama3 8B	100.00± 0.00	100.00± 0.00	100.00± 0.00	100.00± 0.00	100.00± 0.00	100.00± 0.00
5-shot	Vikhr 7B IT 0.4	100.00± 0.00	52.00± 4.00	68.33± 3.33	100.00± 0.00	37.50± 0.00	54.55± 0.00
5-shot MMLU	SaigaLlama3 8B	100.00± 0.00	100.00± 0.00	100.00± 0.00	100.00± 0.00	100.00± 0.00	100.00± 0.00
5-shot MMLU	Vikhr 7B IT 5.4	100.00± 0.00	50.00± 0.00	66.67± 0.00	100.00± 0.00	37.50± 0.00	54.55± 0.00
LoRA	Qwen2 7B IT	100.00± 0.00	58.33± 14.62	72.59± 11.90	100.00± 0.00	41.67± 11.79	57.87± 11.50
SFT	RuBioRoBERTa	100.00± 0.00	55.00± 5.00	70.83± 4.17	100.00± 0.00	27.08± 4.66	42.42± 5.42

Table 20: Results of best models for DSM dataset for various sub-splits for pathology group (mean ± std).

Label	Text (In English)
Pathology	Hello! My name is NAME, I am already AGE years old. I live with my family: my husband and our wonderful daughter. My profession is a midwife in a women's clinic, which I truly love. The team in our department is small, but we are all very close to each other. If any problems arise, we solve them very quickly. Until recently, I was an active person, with a large circle of friends who delighted me with various entertainments: going to the theater, watching movies, walking around the city or just meeting for a cup of coffee. Communicating with them brought me great pleasure. However, recently I began to distance myself from everyone, both at work and in my personal environment. This was the result of my fear, which turned out to be much more powerful than I expected. Fear takes away not only emotional strength, but also physical. At present, it seems to me that it is better to remain alone, avoiding meetings with other people. Although this makes my life less bright, I cannot cope with this condition yet. Also, I used to be into bead embroidery, but now I can't even bring myself to do a simple task. Luckily, we have our cute house cat, who helps me cope with stress and is a great antidepressant. This is my current lifestyle.
Healthy	When I was a child, I accidentally found out that there were people who didn't like me. They were my classmates, and I was very upset when I tried to be friends with them, but my attempt only increased the hostility. I tried to attract their attention with gifts, invitations to visit and other ways, but each new attempt ended in failure and even more trouble. Then I realized that no matter how you behave, there will always be people who don't like you, sometimes for no apparent reason. But it is important to understand that this is normal and you shouldn't try to live up to the expectations of others. You should value yourself for who you are. Many years have passed since then, and now I have many friends and acquaintances, some leave, and new ones come. Each person is unique and beautiful in their own way. I especially liked the expression: "Each person is a small cosmos." It is really profound. Inside each person there is a whole universe consisting of his life experience, mistakes, disappointments, joys, defeats and small victories. If a person allows others to open up, it can be incredibly beautiful and exciting. In a world where we encounter many people every day, I want to believe that everyone respects each other, despite the fact that everyone may not like them. Even those we find unattractive may be dear to someone else. So there is no need to worry about not being loved, because we can find a common language and fill each other with love for life. We are all small universes living together in one big world, trying to get to know each other every day.

Table 21: Paraphrased and anonymized examples from DE dataset.