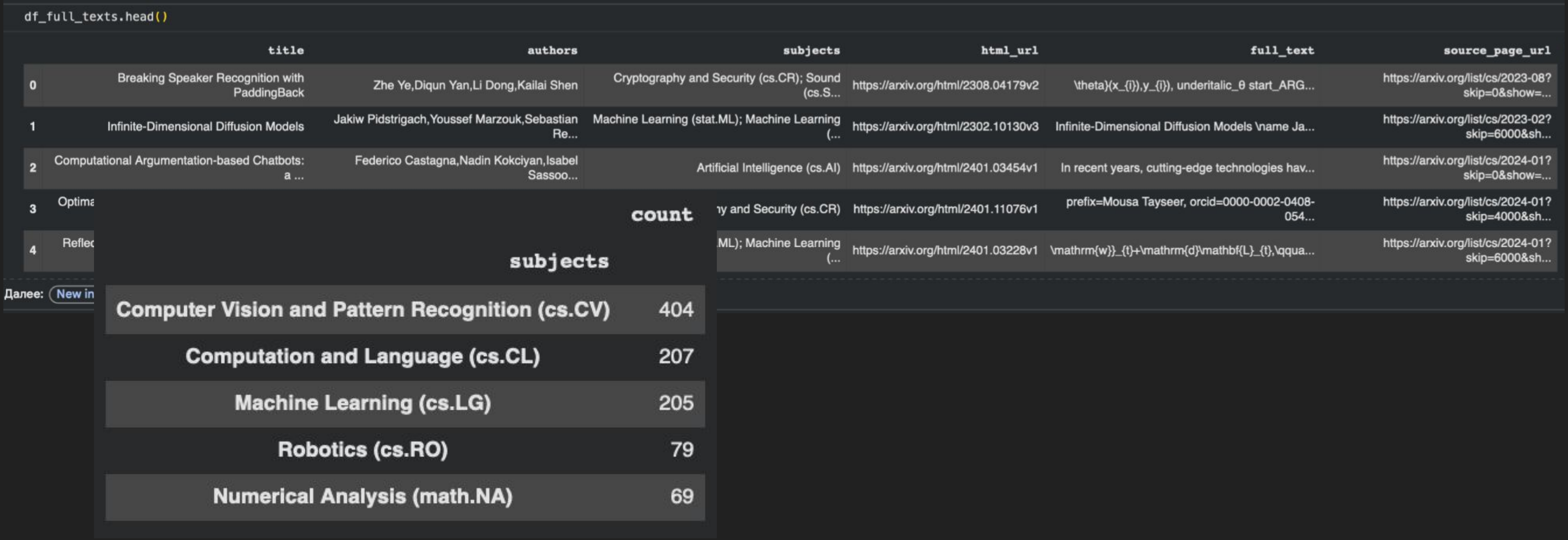
Дообучение моделей BERT (google-bert/bert-base-uncased) и Llama-3.2 (Llama-3.2-3B) для классификации научных статей

Данные

Computer Science

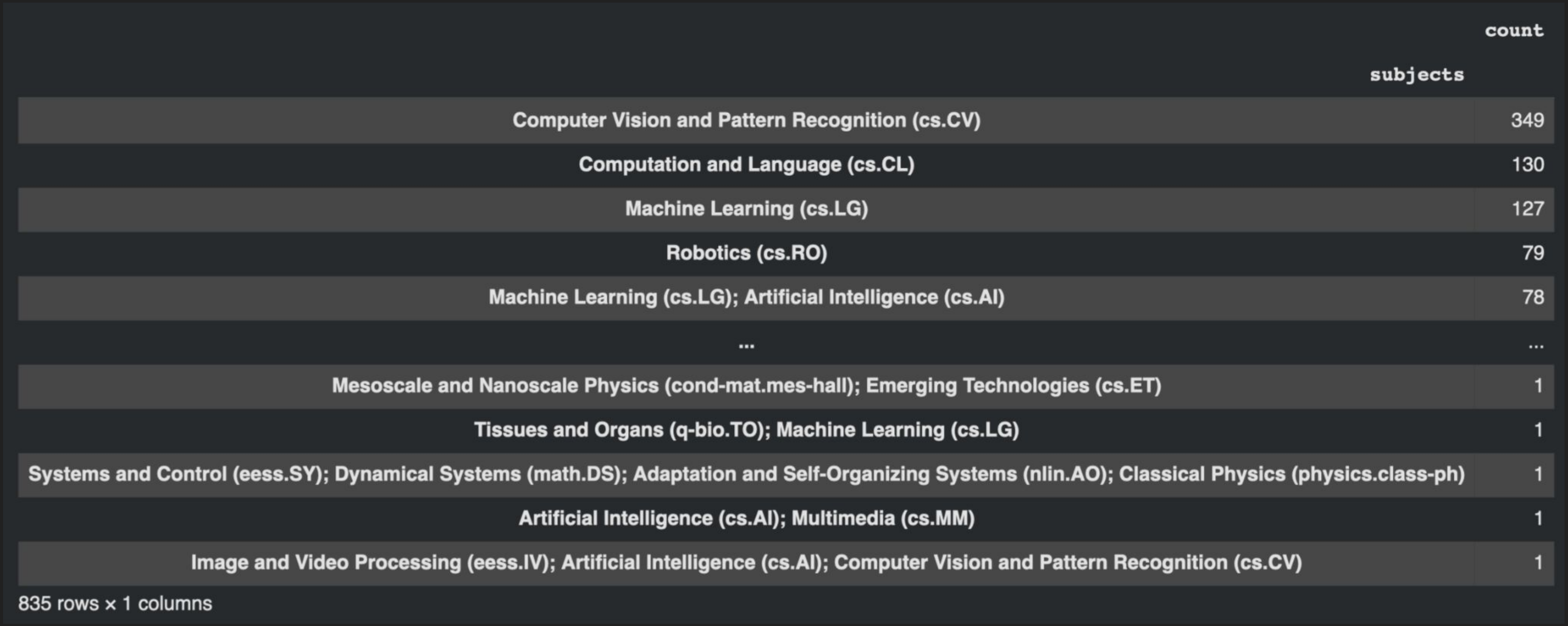


```
df_full_texts.head()
```

	title	authors	subjects	html_url	full_text	source_page_url
0	Breaking Speaker Recognition with PaddingBack	Zhe Ye,Diqun Yan,Li Dong,Kailai Shen	Cryptography and Security (cs.CR); Sound (cs.S...	https://arxiv.org/html/2308.04179v2	\theta}(x_{i}),y_{i}), underitalic_θ start_ARG...	https://arxiv.org/list/cs/2023-08?skip=0&show=...
1	Infinite-Dimensional Diffusion Models	Jakiw Pidstrigach,Youssef Marzouk,Sebastian Re...	Machine Learning (stat.ML); Machine Learning (...	https://arxiv.org/html/2302.10130v3	Infinite-Dimensional Diffusion Models \name Ja...	https://arxiv.org/list/cs/2023-02?skip=6000&sh...
2	Computational Argumentation-based Chatbots: a ...	Federico Castagna,Nadin Kokciyan,Isabel Sassoo...	Artificial Intelligence (cs.AI)	https://arxiv.org/html/2401.03454v1	In recent years, cutting-edge technologies hav...	https://arxiv.org/list/cs/2024-01?skip=0&show=...
3	Optima		ıy and Security (cs.CR)	https://arxiv.org/html/2401.11076v1	prefix=Mousa Tayseer, orcid=0000-0002-0408-054...	https://arxiv.org/list/cs/2024-01?skip=4000&sh...
4	Reflec		ML); Machine Learning (...	https://arxiv.org/html/2401.03228v1	\mathrm{w}}_{t}+\mathrm{d}\mathbf{L}_{t},\qqua...	https://arxiv.org/list/cs/2024-01?skip=6000&sh...

Далее: New in

	count
subjects	
Computer Vision and Pattern Recognition (cs.CV)	404
Computation and Language (cs.CL)	207
Machine Learning (cs.LG)	205
Robotics (cs.RO)	79
Numerical Analysis (math.NA)	69



	count
subjects	
Computer Vision and Pattern Recognition (cs.CV)	349
Computation and Language (cs.CL)	130
Machine Learning (cs.LG)	127
Robotics (cs.RO)	79
Machine Learning (cs.LG); Artificial Intelligence (cs.AI)	78
...	...
Mesoscale and Nanoscale Physics (cond-mat.mes-hall); Emerging Technologies (cs.ET)	1
Tissues and Organs (q-bio.TO); Machine Learning (cs.LG)	1
Systems and Control (eess.SY); Dynamical Systems (math.DS); Adaptation and Self-Organizing Systems (nlin.AO); Classical Physics (physics.class-ph)	1
Artificial Intelligence (cs.AI); Multimedia (cs.MM)	1
Image and Video Processing (eess.IV); Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV)	1

835 rows × 1 columns

Google-bert/bert-base-uncase

```
training_args = TrainingArguments(
  output_dir="./bert-contrastive-lora",
  learning_rate=2e-4,
  per_device_train_batch_size=64,
  num_train_epochs=15,
  eval_strategy="epoch",
  save_strategy="epoch",
  fp16=False,
  load_best_model_at_end=True,
  metric_for_best_model='eval_loss',
  greater_is_better=False
)
```

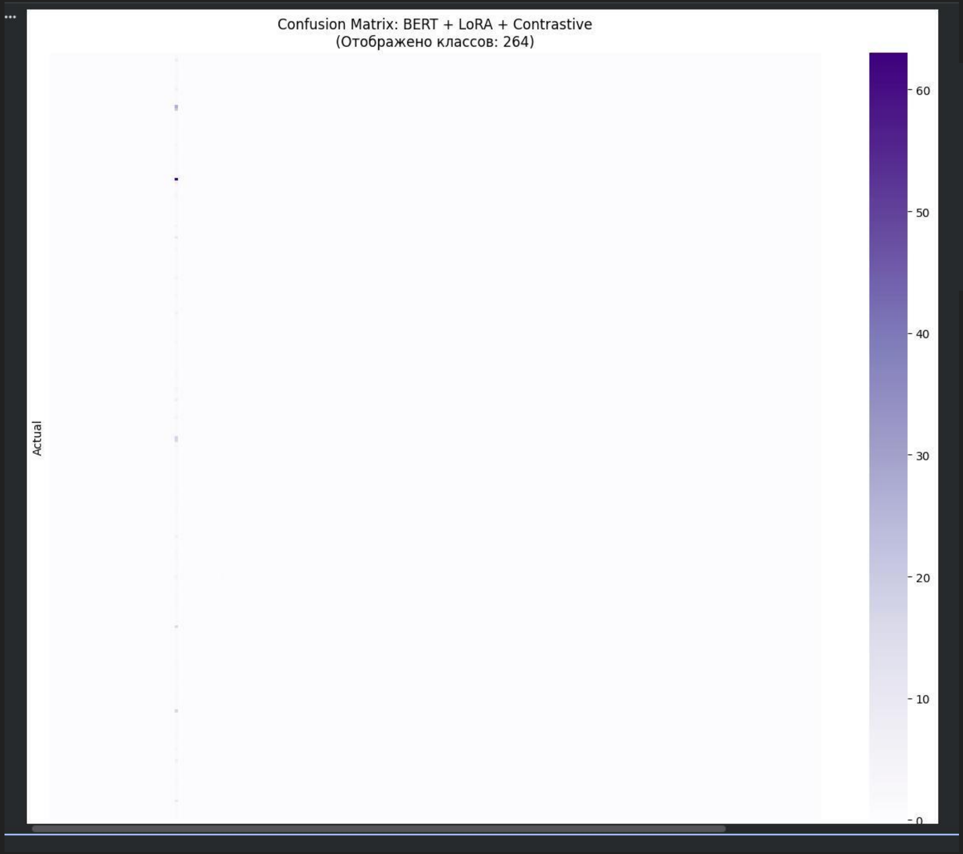
```
trainer = ContrastiveTrainer(
  model=model,
  args=training_args,
  train_dataset=tokenized_ds["train"],
  eval_dataset=tokenized_ds["test"],
  data_collator=data_collator,
  contrastive_alpha=0.05,
  temperature=0.07,
  callbacks=[EarlyStoppingCallback(
      early_stopping_patience=3,
      early_stopping_threshold=0.001
  )]
)
```

unsloth/Llama-3.2-3B

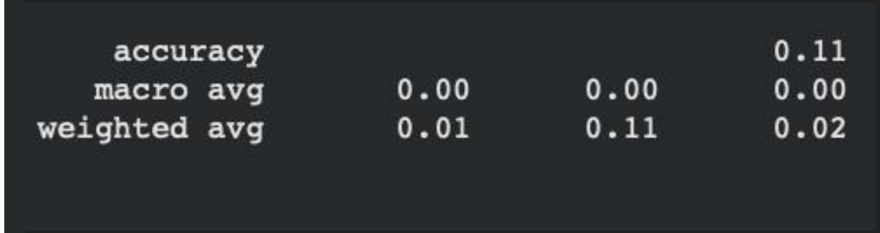
```
model, tokenizer = FastModel.from_pretrained(
   model_name = "unsloth/Llama-3.2-3B",
   max_seq_length = 1024,
   load_in_4bit = True,
   load_in_8bit = False,
   full_finetuning = False
)
```

```
   args = SFTConfig(
       dataset_text_field = "text",
       per_device_train_batch_size = 1,
       gradient_accumulation_steps = 8,
       warmup_steps = 5,
       num_train_epochs = 56,
       learning_rate = 2e-5,
       logging_steps = 1,
    messages = [
        {
            "role": "system",
            "content": "You are a professional academic assistant. Your task is to classify research paper excerpts into their respective arXiv subjects."
        },
        {
            "role": "user",
            "content": f"Analyze this paper excerpt and provide its subject categories:\n\n{truncated_input}."
        },
        {
            "role": "assistant",
            "content": output_text
        },
    ]
```

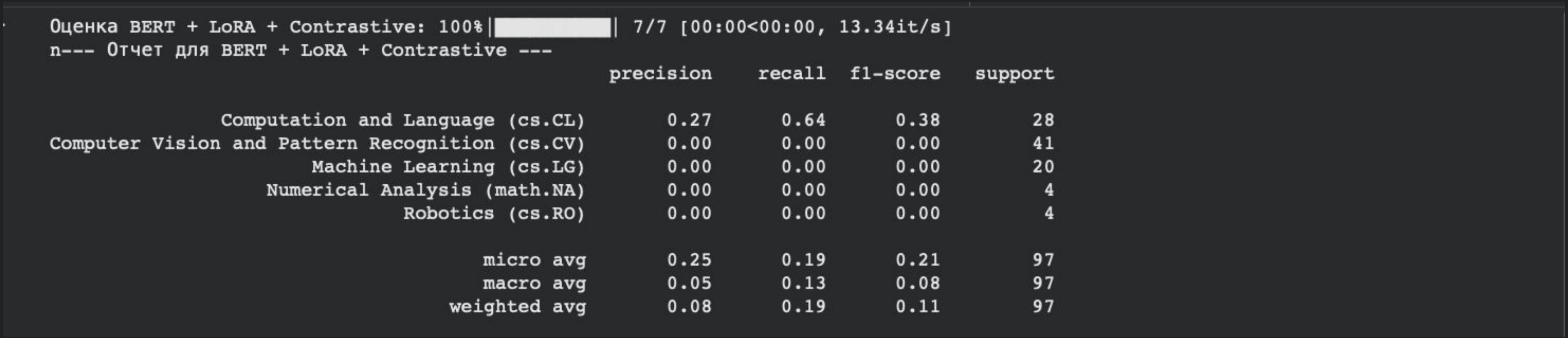
Оценка результатов



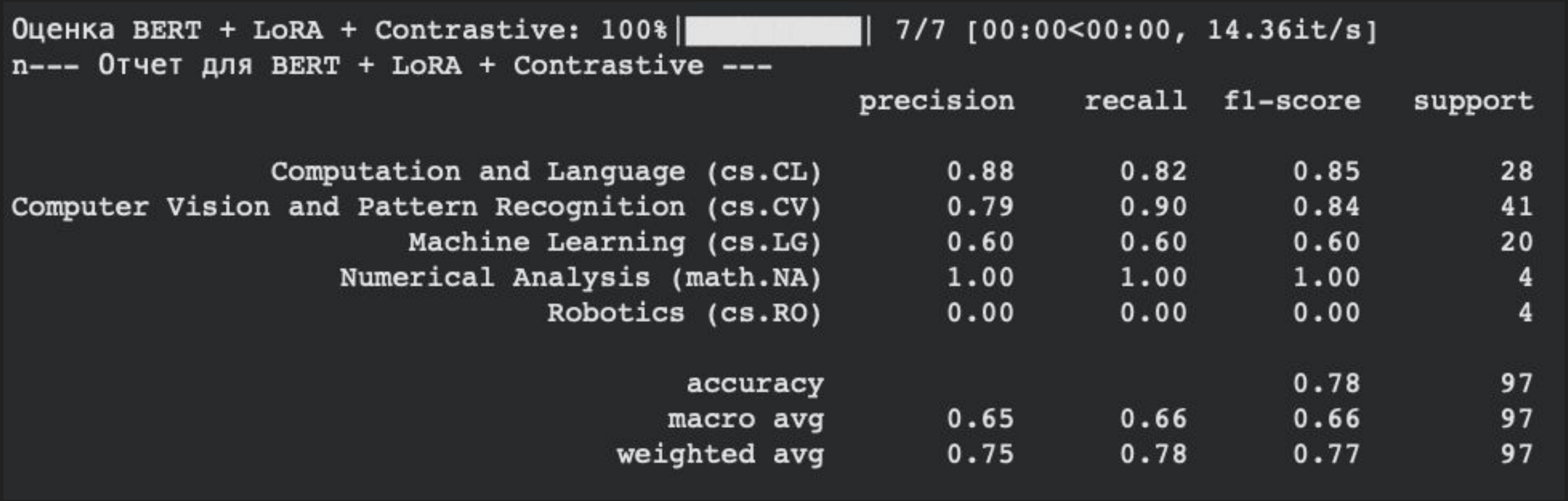
Confusion Matrix: BERT + LoRA + Contrastive
(Отображено классов: 264)



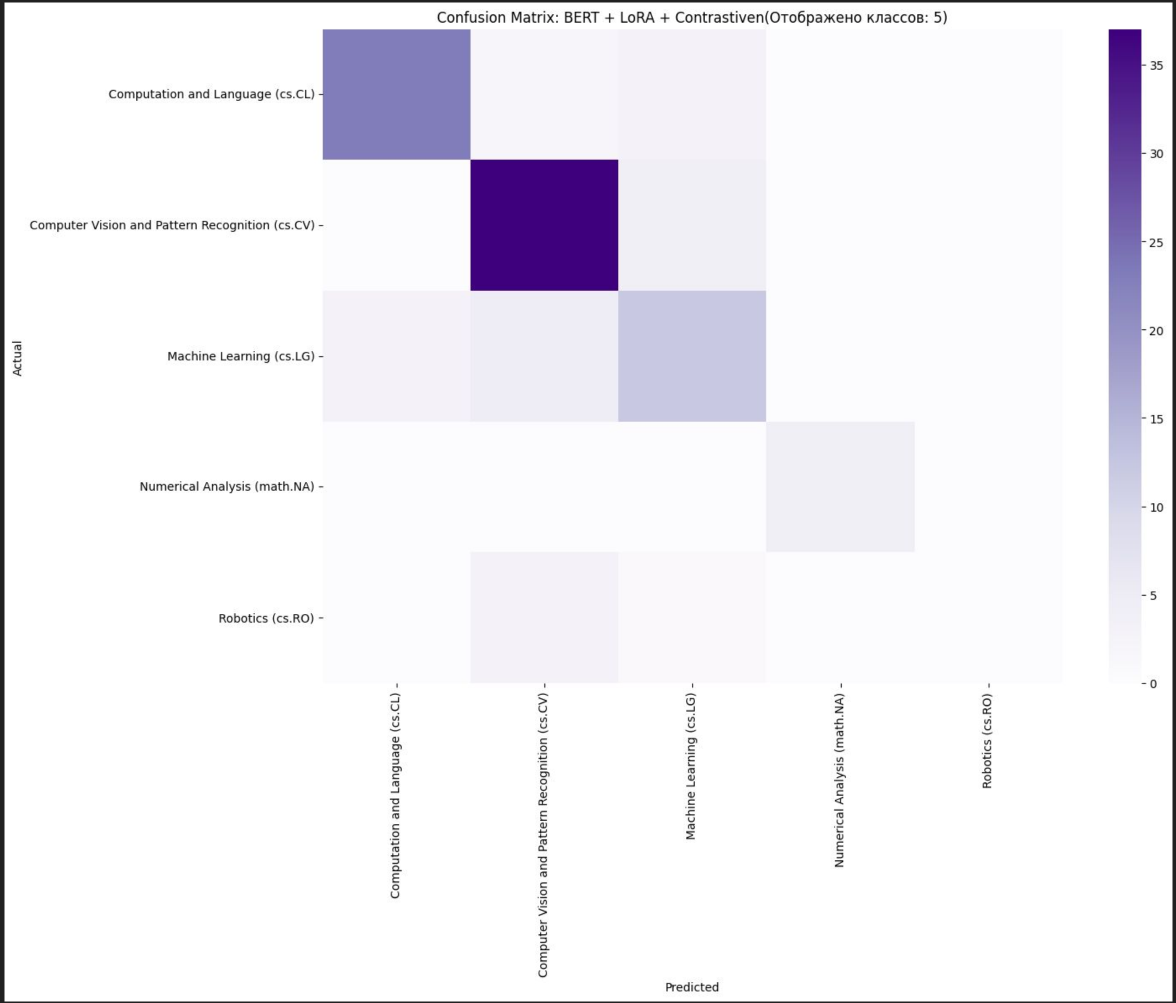
```
    accuracy                           0.11
   macro avg       0.00      0.00      0.00
weighted avg       0.01      0.11      0.02
```



```
Оценка BERT + LoRA + Contrastive: 100%|██████████| 7/7 [00:00<00:00, 13.34it/s]
n--- Отчет для BERT + LoRA + Contrastive ---
                                            precision    recall  f1-score   support

                Computation and Language (cs.CL)       0.27      0.64      0.38        28
Computer Vision and Pattern Recognition (cs.CV)       0.00      0.00      0.00        41
                     Machine Learning (cs.LG)       0.00      0.00      0.00        20
                  Numerical Analysis (math.NA)       0.00      0.00      0.00         4
                             Robotics (cs.RO)       0.00      0.00      0.00         4

                                    micro avg       0.25      0.19      0.21        97
                                    macro avg       0.05      0.13      0.08        97
                                 weighted avg       0.08      0.19      0.11        97
```



```
Оценка BERT + LoRA + Contrastive: 100%|██████████| 7/7 [00:00<00:00, 14.36it/s]
n--- Отчет для BERT + LoRA + Contrastive ---
                                                 precision    recall  f1-score   support

               Computation and Language (cs.CL)       0.88      0.82      0.85        28
Computer Vision and Pattern Recognition (cs.CV)       0.79      0.90      0.84        41
                       Machine Learning (cs.LG)       0.60      0.60      0.60        20
                   Numerical Analysis (math.NA)       1.00      1.00      1.00         4
                               Robotics (cs.RO)       0.00      0.00      0.00         4

                                       accuracy                           0.78        97
                                      macro avg       0.65      0.66      0.66        97
                                   weighted avg       0.75      0.78      0.77        97
```



Confusion Matrix: BERT + LoRA + Contrastiven(Отображено классов: 5)

Предсказанная категория (Llama-3.2-3B):

The following is a list of the 10 most common subjects in the arXiv.org database. The list is ordered by the number of papers in each subject. The list is based on the 2018-12-01 snapshot of the arXiv database. The list is not exhaustive and is subject to change. The list is not a recommendation for which subjects to study. It is a list of the most popular subjects in the arXiv database. The list is not a recommendation for which subjects to study. It is a list of the most popular subjects in the arXiv database.

The histogram is a fundamental tool for summarizing and analyzing large data sets. It is a binned representation of the distribution of a continuous variable, where the bins are typically of equal width and the counts of observations falling within each bin are recorded. The histogram is a compact representation of the data, and its properties can be used to make inferences about the data. For example, the mean of a histogram is the center of mass of the bins, and the variance is the sum of the squares of the distances of each bin from the mean. The histogram is also a useful tool for comparing the distributions of different data sets.