RoQLlama: A Lightweight Romanian Adapted Language Model

George-Andrei Dima[1], **Andrei-Marius Avram**[1], **Cristian-George Crăciun**[1,2], **Dumitru-Clementin Cercel**[*][1]

[1]National University of Science and Technology POLITEHNICA Bucharest, Romania
[2]Technical University Munich, Munich, Germany

Abstract

The remarkable achievements obtained by open-source large language models (LLMs) in recent years have predominantly been concentrated on tasks involving the English language. In this paper, we aim to advance the performance of Llama2 models on Romanian tasks. We tackle the problem of reduced computing resources by using QLoRA for training. We release RoQLlama-7b, a quantized LLM, which shows equal or improved results compared to its full-sized counterpart when tested on seven Romanian downstream tasks in the zero-shot setup. Also, it consistently achieves higher average scores across all few-shot prompts. Additionally, we introduce a novel Romanian dataset, namely RoMedQA, which contains single-choice medical questions in Romanian.

1 Introduction

Transformer models (Vaswani et al., 2017) represent the state-of-the-art solution adopted by natural language processing (NLP) tasks (Wilie et al., 2020). Due to current breakthroughs in computational capabilities, models were scaled in terms of parameters, acquiring new remarkable abilities in terms of natural language understanding (Elmadany et al., 2023). As a result, a series of proprietary and open large language models (LLMs) were created.

One major downside of LLMs is the enormous amount of computational resources and training data they require. Democratizing LLMs constitutes a vital research direction, increasing the possibility of breakthroughs. In this sense, we release a new lightweight Romanian language-adapted LLM with 7 billion parameters and quantized to 4 bits by employing the state-of-the-art quantized LoRA (QLoRA) training technique (Dettmers et al., 2024). We evaluate our model on several Romanian datasets, covering seven tasks and comparing it to its original counterpart. Our results showed that RoQLlama-7b outperformed the other Llama models on four out of the seven tasks investigated using zero-shot prompting. Furthermore, due to quantization, the model has a significantly smaller memory footprint, up to three times less than the base model.

To summarize, the contributions of this work are:

- We train and release the first Romanian-adapted LLM based on Llama2-7b (Touvron et al., 2023b), with reduced memory footprint, called **RoQLlama-7b**[1].
- We introduce **RoMedQA**[2], the first dataset comprising medical exam questions in the Romanian language.
- We comprehensively test the released model, comparing it with the Llama2-7b models on Romanian downstream tasks.
- We investigate parameter efficiency and language adaptation in a low-resource language setting.

2 RoQLlama

2.1 Training Dataset

When building our training data, we start from the work done by Masala et al. (2020). We use RoWiki, a Romanian Wikipedia dump containing 0.3 GB of text, and RoTex, a text collection from online Romanian sources containing 1.5 GB of text.

Also, we included in our training data the Romanian sections from the OSCAR corpus (Suárez et al., 2019), containing 45.6 GB of text, and from

[*] Corresponding author: dumitru.cercel@upb.ro.

[1] https://huggingface.co/andreidima/Llama-2-7b-Romanian-qlora

[2] https://huggingface.co/datasets/craciuncg/RoMedQA_v1

Findings of the Association for Computational Linguistics: EACL 2024, pages 4531–4541
November 12-16, 2024 ©2024 Association for Computational Linguistics

Model	M1 (GB)	M2 (GB)
Llama2-7b	13.4	14.8
RoQLlama-7b	4.7	6.1

Table 1: Memory footprints of Llama2-7b and RoQLlama-7b. M1 represents the VRAM used by the model, whereas M2 represents the VRAM needed to ingest a prompt of 1,000 tokens.

the CC-100 corpus (Conneau et al., 2020), containing 61.4 GB of text. Touvron et al. (2023a) suggest that various pre-processed CommonCrawl (Smith et al., 2013) variants could enhance the obtained results. Therefore, we decided to use the Romanian corpora from both OSCAR and CC-100 in our training data, even though both are based on the same data source. See Appendix C for information on the steps involved in processing the training dataset.

2.2 Training Process

We trained the model using QLoRA. We take the advice from (Dettmers et al., 2024) regarding the low-rank adaptation (LoRA) (Hu et al., 2021) adapter hyperparameters, as they believe that if LoRA is applied to every layer, *LoRA r* will not impact the experimental results. We applied LoRA to all linear layers of Llama2 and set *LoRA r* at 8. We also kept *LoRA alpha* at 8 and set *LoRA dropout* at 0.05 since dropout has been shown to boost performance in the smaller Llama variants (Dettmers et al., 2024). Regarding the quantization of the base models, we used the 4-bit NF4 quantization and did not apply double quantization.

We used the paged AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 1e-5, a weight decay of 0.001, and a gradient clipping set at 0.01. In order to fit the graphics processing unit (GPU), we use a per-device batch size of 2 with 4 gradient accumulation steps, resulting in a batch size of 8. The model was trained for 900,000 steps, which is approximately 7.3B tokens.

2.3 Memory Footprint

In Table 1, we compare our quantized model with the original version regarding memory footprint and processing time. M1 represents the video random-access memory (VRAM) usage measured after loading each model into the GPU, with the original model loaded in float16. M2 indicates the maximum VRAM used when processing an artificial prompt of 1,000 tokens. All tests were conducted on an A100 80GB NVIDIA GPU.

Our model has a significantly smaller memory footprint, reducing both the M1 and M2 memory required to run it from 13.4 GB to 4.7 GB and from 14.8 GB to 6.1 GB, respectively.

3 RoMedQA

More publicly available datasets are needed for the Romanian NLP tasks. To contribute to this area, we introduce **RoMedQA**, a dataset that amounts to 4,127 single-choice questions regarding the medical field in the Romanian language. The dataset consists of advanced biology questions used in entrance examinations in medical schools in Romania. Each question has five possible answer choices, numbered from 1 to 5, with only one correct answer. See Appendix B for a more detailed dataset description.

4 Evaluation

We evaluate the RoQLlama-7b model and the original Llama2 on seven Romanian NLP tasks as follows:

- **Medical Question Answering** - using the new dataset (RoMedQA) introduced in this work.
- **Question Answering** - using the Romanian subset (RoQA) (Dumitrescu et al., 2021) of the Cross-Lingual Question Answering Dataset (xSQuAD) (Artetxe et al., 2020).
- **Emotion Detection** - using the second version of Romanian Emotion Dataset (REDv2) (Ciobotaru et al., 2022).
- **Romanian/Moldavian Dialect Classification** - using the Moldavian and Romanian dialectal Corpus (MOROCO) (Butnaru and Ionescu, 2019).
- **Satire Detection** - using the Satire detection Romanian Corpus (SaRoCo) (Rogoz et al., 2021).
- **News Summarization** - on the Romanian Summarization dataset (RoSum) (Niculescu et al., 2022).
- **Textual Similarity** - using the Romanian Semantic Textual Similarity dataset (RoSTS) (Dumitrescu et al., 2021)

Model	RoMedQA	RoQA	REDv2	RoMD	SaRoCo	RoSum	RoSTS	Avg.
Llama2-7b	3.60	24.88	3.59	4.95	28.17	18.47	-0.663	14.36
Llama2-7b-chat	1.79	**44.05**	**6.89**	20.38	**29.88**	22.26	0.039	25.31
RoQLlama-7b	**3.67**	39.64	6.45	**29.78**	29.63	**19.46**	**0.401**	**28.38**

Table 2: Zero-shot results of the Llama2 models on all the 7 evaluated tasks, together with the average score.

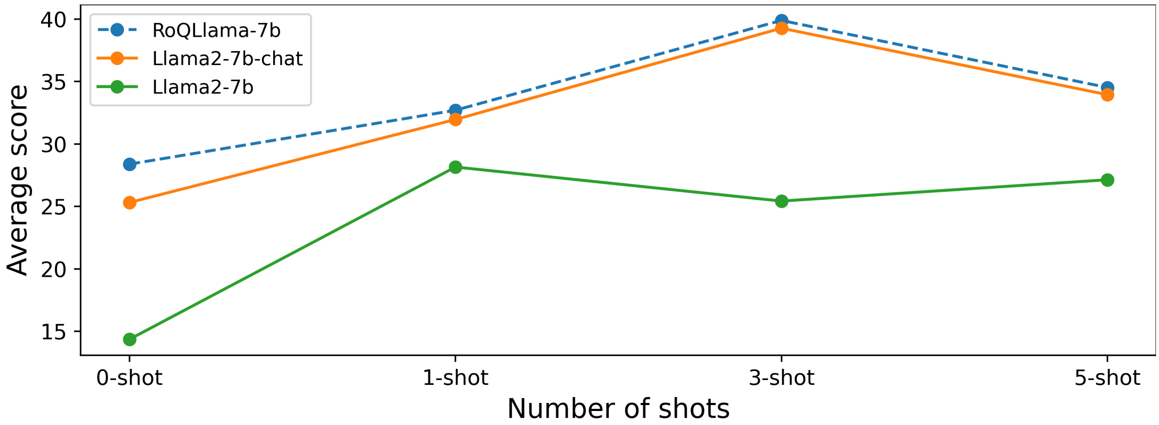

Figure 1: Average few-shot results of the Llama2 models.

We compare them to the original baselines, which include Ro-BERT (Dumitrescu et al., 2020), Ro-GPT2 (Niculescu et al., 2021), and various other architectures. During the evaluation process, we kept the same configuration for all models and all tasks: *temperature* = 0.6, and *top_p* = 0.9. Also, we stopped the generation at the first new line. We varied the maximum number of generated tokens for different task categories: 10 for classification and regression tasks (RoMedQA, REDv2, RoMD, SaRoCo, RoSTS), 250 for question answering (RoQA), and 2,048 for summarization (RoSum).

We show the original prompts used to evaluate the language models on each task, as well as their translation in English in Appendix D.

5 Results

The results using zero-shot prompting are depicted in Table 2, which outlines the macro F1-scores on each classification task (i.e., RoMedQA, REDv2, RoMD, and SaRoCo), the overlap F1-score on the question answering task (i.e., RoQA), the ROUGE-L score on the summarisation task (i.e., RoSum), and the Pearson correlation score for the regression task (i.e., RoSTS). The RoQLlama-7b model obtained the highest score on four out of the seven evaluated tasks, namely on RoMedQA, RoMD, RoSum, and RoSTS. The highest average score of 28.39 was obtained by Llama2-7b, outperforming Llama2-7b-chat by ∼3% and almost doubling the average score[3].

We also evaluate the performance of the newly introduced RoQLlama-7b model on few-shot prompting (i.e., one-shot, three-shot, and five-shot). This allows us to analyze how the model improves when additional examples for each task are given. The results of this analysis are depicted in Figure 1, which outlines the variations in average scores when an increasing number of examples are presented to each model in the prompt. We can observe that RoQLlama-7b consistently outperforms both Llama2-7b and Llama2-7b-chat on all the few shots prompting tasks evaluated in this work.

Furthermore, because the language models can hallucinate and produce inadequate output for each classification task, we also evaluate the *not followed instruction* (NFI) score, which measures the percentage of samples on which the model fails to adhere to the given instructions. The following subsections present results on zero-shot prompting for each evaluation task.

5.1 RoMedQA

To establish a baseline for comparing these results with other future results that can be achieved by

[3] To compute the average score, we normalized the Pearson score which we obtained on RoSTS.

Model	Acc	F1	NFI
Llama2-7b	**22.77**	3.60	0.00
Llama2-7b-chat	11.69	1.79	0.00
RoQLlama-7b	21.57	**3.67**	0.00

Table 3: Evaluation results of our models on the RoMedQA dataset. The scores of Llama2 models are calculated using zero-shot prompting.

Model	EM	F1
mBERT	58.99	72.69
XLM-R Large	69.66	83.56
Llama2-7b	13.94	24.88
Llama2-7b-chat	**25.12**	**44.05**
RoQLlama-7b	24.20	39.64

Table 4: Evaluation results of our models on the Romanian xSQuAD subset. The scores of Llama2 models are calculated using zero-shot prompting. Non-Llama baselines are taken from (Dumitrescu et al., 2021).

training, we evaluate the test split of the RoMedQA, which contains 831 entries.

As shown in Table 3, RoQLlama-7b outperforms both Llama2-7b models regarding macro F1-score. However, the scores of all three models remain unsatisfactory. This is expected, as none of the models are trained explicitly on medical data, which is essential for solving the questions in the dataset.

5.2 RoQA

We evaluate the original and the Romanian Llama2 models on the RoQA, which contains 240 paragraphs and 1,190 question-answer pairs annotated using the SQuAD v1.1 guidelines (Rajpurkar et al., 2016). Table 4 depicts the results with zero-shot prompting. With an Exact Match (EM) score of 24.20 and an overlap F1-score of 39.64, RoQLlama-7b performs better than its original counterpart in terms of both EM and overlap F1-score, outperforming the Llama2-7b model by a considerable margin. However, its performance is slightly worse than that of the Llama2-7b-chat.

5.3 REDv2

We evaluate the REDv2 dataset, which contains collected tweets in the Romanian language, annotated with their associated emotions. RoQLlama-7b performs better than Llama2-7b in terms of macro F1-score and accuracy. However, the highest accuracy is obtained by the Llama2-7b-chat model, with 48.24%, almost double that of the other two models. Llama2-7b-chat also obtains the highest macro F1-score on REDv2.

Model	Acc	F1	NFI
Ro-BERT	54.1	66.8	-
XLM-RoBERTa	50.4	61.9	-
Llama2-7b	26.46	3.59	0.00
Llama2-7b-chat	**48.24**	**6.89**	0.00
RoQLlama-7b	26.34	6.45	0.00

Table 5: Evaluation results of our models on the REDv2 dataset. The scores of Llama2 models are calculated using zero-shot prompting. Non-Llama baselines are taken from (Ciobotaru et al., 2022).

Model	Acc	F1	NFI
KRR + $k_6^{0/1}$	94.13	94.06	-
CNN	92.75	92.71	-
CNN + SE	92.99	92.93	-
Llama2-7b	4.42	4.95	91.36
Llama2-7b-chat	30.58	20.38	42.01
RoQLlama-7b	**46.84**	**29.78**	**11.71**

Table 6: Evaluation results of our models on the RoMD dataset. The scores of Llama2 models are calculated using zero-shot prompting. Non-Llama baselines are taken from (Butnaru and Ionescu, 2019).

5.4 RoMD

We evaluate the Romanian Llama2 model introduced in this work on a classification task to determine whether a text belongs to the Romanian or Moldavian dialects using the test subset of the MOROCO. The results are depicted in Figure 6. RoQLlama-7b obtains the highest accuracy and macro F1-score out of all Llama2 models, with a 46.84% accuracy and a 29.78% macro F1-score. Also, it shows the lowest NFI score, the model not following instructions in 11.71% of the cases.

5.5 SaRoCo

The SaRoCo introduces a dataset designed for identifying satirical content in Romanian news articles (Rogoz et al., 2021). The results depicted in Table 7 outline that Llama2-7b-chat outperformed RoQLlama-7b both in terms of accuracy, where it achieved a score of 50.12% and macro F1-score

Model	Acc	F1	NFI
Ro-BERT	73.00	71.50	-
Char-CNN	69.66	71.09	-
Llama2-7b	11.69	28.17	21.27
Llama2-7b-chat	**50.12**	**29.88**	8.72
RoQLlama-7b	41.12	29.63	**7.98**

Table 7: Evaluation results of our models on the SaRoCo dataset. The scores of Llama2 models are calculated using zero-shot prompting. Non-Llama baselines are taken from (Butnaru and Ionescu, 2019).

Model	R-1	R-2	R-L
Ro-GPT2-base	34.80	19.91	34.16
Ro-GPT2-medium	35.46	20.61	34.67
Ro-GPT2-large	34.92	19.95	33.84
Llama2-7b	24.67	12.22	18.47
Llama2-7b-chat	**32.05**	**14.72**	**22.26**
RoQLlama-7b	24.37	11.70	18.46

Table 8: Evaluation results of our models on the Romanian summarization dataset. The scores of Llama2 models are calculated using zero-shot prompting. Non-Llama baselines are taken from (Niculescu et al., 2022).

with 29.88%. However, RoQLlama-7b obtained the lowest NFI score out of all three models, with 7.98% of the answers being inadequate concerning the provided instructions.

5.6 RoSum

We compare RoQLlama-7b with the original Llama2 models on RoSum for summarization performance. This summarization dataset was created by crawling Romanian news articles, which also provided bullet point summaries. Table 8 presents the results, which indicate that Llama2-7b-chat and, to a lesser extent, Llama2-7b outperform our model. This may be due to our model's training being conducted on relatively small samples of Romanian text, each with fewer than 1,024 tokens.

5.7 RoSTS

The original and Romanian Llama2 models were also evaluated on their performance for textual similarity using the test set of the RoSTS dataset, which contains 1,379 sentence pairs, each annotated with a similarity score from 0 to 5. We compare the models by computing both the Pearson and Spearman correlation coefficients. The results are depicted in Table 9. RoQLlama-7b performs significantly better on this task than both Llama2-7b and Llama-7b-chat, with the highest Pearson correlation (0.412) and the highest Spearman correlation (0.462).

Model	Pearson	Spearman
RNN	0.685	-
mBERT (cased)	0.766	-
mBERT (uncased)	0.769	-
Ro-BERT (cased)	0.792	-
Ro-BERT (uncased)	0.815	-
Llama2-7b	-0.663	-0.541
Llama2-7b-chat	0.039	0.055
RoQLlama-7b	**0.401**	**0.462**

Table 9: Evaluation results of our models on the RoSTS test set. The scores of Llama2 models are calculated using zero-shot prompting. Non-Llama baselines are taken from (Dumitrescu et al., 2021).

6 Conclusions

In this paper, we advance state-of-the-art NLP techniques for Romanian and address the scarcity of Romanian datasets. We introduce a lightweight LLM for Romanian and a new medical dataset of single-choice exam questions. RoQLlama-7b, a quantized version of Llama2-7b, achieves higher average scores across Romanian tasks while using three times less memory.

RoMedQA is the first Romanian dataset of medical questions and answers based on entrance exams for medical schools in Romanian. It is valuable for training and testing LLMs in medical knowledge and language comprehension. Future work includes enhancing the dataset with contextual information, adapting it for smaller models, and integrating the results into the Romanian LiRo benchmark (Dumitrescu et al., 2021).

Limitations

Since our model is a fine-tuned version of the Llama2 model, it inherits the existing limitations of the parent model, as shown by Touvron et al. (2023b). Additionally, our model was further trained on Internet text so that it may have been exposed to specific biases prevalent on the Romanian Internet. Users should be aware that this model carries risks of generating hallucinations, toxic language, and various biases.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

Andrei Butnaru and Radu Tudor Ionescu. 2019. Moroco: The moldavian and romanian dialectal corpus. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 688–698.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Alexandra Ciobotaru, Mihai Vlad Constantinescu, Liviu P Dinu, and Stefan Dumitrescu. 2022. Red v2: enhancing red dataset for multi-label emotion detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1392–1399.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of romanian bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, et al. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Abdelrahim Elmadany, Elmoatez Billah Nagoudi, and Muhammad Abdul Mageed. 2023. Orca: A challenging benchmark for arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586.

Rintaro Enomoto, Arseny Tolmachev, Takuro Niitsuma, Shuhei Kurita, and Daisuke Kawahara. 2024. Investigating web corpus filtering methods for language model development in japanese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 154–160.

Pablo Gamallo, Pablo Rodríguez, Iria de Dios-Flores, Susana Sotelo, Silvia Paniagua, Daniel Bardanca, José Ramom Pichel, and Marcos Garcia. 2024. Open generative large language models for galician. *arXiv preprint arXiv:2406.13893*.

Aryo Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. 2023. Parameter-efficient fine-tuning of llama for the clinical domain. *arXiv preprint arXiv:2307.03042*.

Joseph Gesnouin, Yannis Tannier, Christophe Gomes Da Silva, Hatim Tapory, Camille Brier, Hugo Simon, Raphael Rozenberg, Hermann Woehrel, Mehdi El Yakaabi, Thomas Binder, et al. 2024. Llamandement: Large language models for summarization of french legislative proposals. *arXiv preprint arXiv:2401.16182*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zixian Huang, Yulin Shen, Xiao Li, Gong Cheng, Lin Zhou, Xinyu Dai, Yuzhong Qu, et al. 2019. Geosqa: A benchmark for scenario-based question answering in the geography domain at high school level. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5866–5871.

Omri Keren and Omer Levy. 2021. Parashoot: A hebrew question answering dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 106–112.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. Robert–a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637.

Quan Nguyen, Huy Pham, and Dung Dao. 2023. Vinallama: Llama-based vietnamese foundation model. *arXiv preprint arXiv:2312.11011*.

Mihai Alexandru Niculescu, Stefan Ruseti, and Mihai Dascalu. 2021. Rogpt2: Romanian gpt2 for text generation. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1154–1161. IEEE.

Mihai Alexandru Niculescu, Stefan Ruseti, and Mihai Dascalu. 2022. Rosummary: Control tokens for romanian news summarization. *Algorithms*, 15(12):472.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.

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

Ana-Cristina Rogoz, Gaman Mihaela, and Radu Tudor Ionescu. 2021. Saroco: Detecting satire in a novel romanian corpus of news articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1073–1079.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: An italian instruction-tuned llama. *arXiv preprint arXiv:2307.16456*.

Jason R Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: a large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3025–3035.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 33469–33484.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeffrey Ward. 2024. A continued pretrained llm approach for automatic medical note generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 565–571.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024. Tablellama: Towards open large generalist models for tables. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6024–6044.

A Related Work

English Llama Models. Llama (Touvron et al., 2023a) is a family of large language models trained exclusively on publicly available data and released openly, ranging in sizes from 7 to 65 billion parameters and with a context window of 2048 tokens. Llama2 (Touvron et al., 2023b) was introduced as a collection of large language models which showed improved performance compared to the first generation. Llama2 models have sizes ranging from 7 to 70 billion parameters and a context window of 4096 tokens. The versatility and accessibility of the models from the Llama family render them some of the most popular large language models.

The first-generation and second-generation Llama models were the foundation for numerous research papers. For instance, Xie et al. (2023) fine-tuned Llama1 on an instruction dataset containing various tasks from the financial domain, showing the potential for domain-specific tuning of large language models. Yuan et al. (2024) further trained Llama2-13b on texts from the medical domain in order to approach the problem of medical text generation.

Llama Models in Other Languages. Adapting Llama models to low-resource languages has significantly improved downstream task performance for those languages. Pires et al. (2023) further trained Llama1 (sizes 7B and 65B) on a Portuguese corpus containing 7.3 billion tokens, resulting in better performance on Portuguese tasks. Similarly, Cui et al. (2023) trained Llama 7B, 13B, and 33B on a Chinese corpus, achieving superior results in Chinese text comprehension and generation compared to the original model. Kuulmets et al. (2024) adapted Llama2 on Estonian while maintaining its performance in English by training on a combined English and Estonian corpus of 5 billion tokens. Also, many other languages have benefited from adapted versions of Llama, such as French (Gesnouin et al., 2024), Italian (Santilli and Rodolà, 2023; Basile et al., 2023), Japanese (Enomoto et al., 2024), Galician (Gamallo et al., 2024), and Vietnamese (Nguyen et al., 2023).

QLoRA. Parameter efficient fine-tuning (PEFT) techniques (Mangrulkar et al., 2022; Xu et al., 2023) aim to adapt large models for specific tasks with minimal computational resources, addressing the challenges posed by their enormous scale. A well-known PEFT method is LoRA (Hu et al., 2021), which reduces the resources needed to train large language models by training only the rank-decomposition matrices corresponding to the dense layers of the Transformer architecture instead of full parameter training. Models trained with LoRA have been shown to require up to three times less memory than usual training with very little to almost no loss in performance (Hu et al., 2021).

QLoRA (Dettmers et al., 2024) further reduces the required resources by introducing 4-bit normal float quantization (NF4), double quantization (DQ), and paged optimizers. NF4 is a novel quantization technique that goes beyond a crude model approximation and cleverly uses the available 4 bits to minimize the loss of information in the model's parameters. Using the normal distribution, NF4 eliminates outlier parameters and accurately represents the more often occurring parameter values. DQ further achieves memory savings by quantizing the quantization constants. The memory overhead caused by quantization constants is typically 0.5 bits per parameter. When using DQ, quantization constants have a memory footprint of 0.127 bits per parameter.

Building on the advantages of PEFT training, numerous research papers have focused on fine-tuning Llama models with LoRA or QLoRA. Gema et al. (2023) fine-tuned Llama for the clinical domain using LoRA and reported state-of-the-art results across clinical tasks. Zhang et al. (2024) addressed tasks based on semi-structured tables by building TableLama, a Llama2-7b model fine-tuned with LongLoRA (Chen et al., 2023). Santilli and Rodolà (2023) used LoRA to fine-tune Llama on a corpus of instructions translated into Italian and reported competitive results on Italian downstream tasks. Basile et al. (2023) built LLaMAntino by adapting the Llama2-7b and 13B models to the Italian language using QLoRA.

B RoMedQA

Overview. One of the significant issues in the Romanian NLP tasks is the need for more available data. We decided to make our contribution based on several works (Wang et al., 2022; Keren and Levy, 2021; Huang et al., 2019) performed by various linguistic research communities that enriched data availability in their respective languages, opening the doors to new research possibilities in terms of intra and inter-lingual NLP. In light of the above, we introduce **RoMedQA**, a dataset that amounts to

Word	Translation	TF-IDF Score
celulă	cell	0.02205
mușchi	muscle	0.02192
nerv	nerve	0.02026
nivel	level	0.01998
corect	correct	0.01991
răspuns	answer	0.01935
niciun	none	0.01895
afirmație	statement	0.01848
fibră	fiber	0.01810
următor	next	0.01754

Table 10: TF-IDF scores of the most common words found in RoMedQA.

4,127 single-choice questions regarding the medical field in the Romanian language. The dataset consists of advanced biology questions used in entrance examinations in medical schools in Romania.

Data Collection. Building this dataset was quite challenging because of the variety in the data format. We had to resort to multiple techniques to collect the entries. The questions given at past entrance examinations were available on the medical universities' websites in HTML format, PDF documents, or scans. Where possible, we used web scrapping to extract the questions with their respective correct answer. In other cases, we had to scrape PDF documents for text or, less favorably, perform OCR on the PDF scans to extract the underlying questions and answers.

Unfortunately, some scans were not of good quality. Therefore, we manually extracted the questions written on the scanned documents. Ultimately, we manually inspected the data to identify and rectify any noise introduced by the OCR process. This ensures that our dataset is of good quality, with no issues for anyone to use, eliminating the need for sanitization and other data pre-processing.

Data Analysis. Each question has five possible answer choices, numbered from 1 to 5, with only one correct answer. We can notice that the classes are pretty balanced, as depicted in Figure 2, ensuring that the dataset can be used for potential training and relevant results can be achieved through testing. We first remove stop-words and lemmatize the given entries to compute the TF-IDF scores (Sparck Jones, 1972) in Table 10. We compute the TF-IDF score by multiplying the absolute frequency of each word in the corpus by the logarithm of the IDF. This gives us insights into the dataset's most common subtopics of biology. In Figure 3, we show the token length distribution of the dataset. The tokens were computed using the Llama tokenizer.

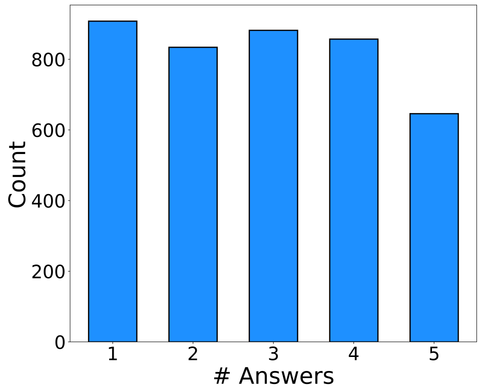

Figure 2: Class distribution of the RoMedQA dataset.

C RoQLlama Training Data Processing

The dataset used for training RoQLlama contained a significant amount of noise, with samples in Slavic languages being incorrectly labeled as Romanian. Given the large amount of data compared to our limited computing resources, we choose a greedy approach to clean it. We split each text into sentences using the NLTK sentence tokenizer[4] and then removed any sentences that included characters not based on Latin characters. This step helped eliminate most of the mislabeled text. After that, we combined the cleaned sentences into text samples with fewer than 1,024 tokens for training the model.

D Evaluation Prompts

In this section, we include the prompts used for testing the Llama2 models, as well as our newly introduced RoQLlama model. This section covers the translated prompts and the prompts used in the Romanian language. These are the prompts used for zero-shot inference. For the few-shot setting, multiple examples are given, one below another, along with the answer keys and format, in the same way the model should respond in the zero-shot scenario.

[4] https://www.nltk.org/api/nltk.tokenize.html

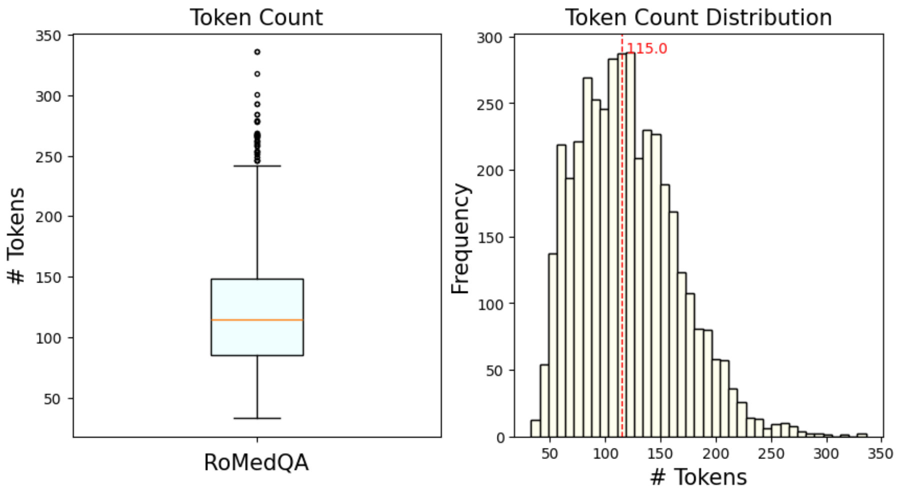

Figure 3: An overview of the sample length distribution regarding the number of tokens.

D.1 Translated Prompts in English

RoMedQA

I answer medical multiple-choice questions using only the digit of the correct answer from the answer choices. There is only one correct answer.

Question: {}
Answer:

RoQA

I read the given context and briefly answer the given questions, using only information from the context. I do not offer any explanation.

Context: {}
Question: {}
Answer:

REDv2

I read the following text and annotate it based on its predominant emotion.

The only emotion categories I can choose from are: Sadness, Surprise, Fear, Anger, Neutral, Trust, and Joy. I do not offer any explanation.

Text: {}
Emotion:

RoSTS

I read both sentences and semantically annotate their similarity with a score, scoring them from 0 (the sentences have no semantic similarity) to 1 (the sentences are identical semantically). I do not offer any explanation.

Sentence1: {}
Sentence2: {}
Semantic similarity score:

RoMD

I read the following paragraph and annotate it based on its dialect, Romanian or Moldavian.

The only dialect categories I can choose from are Romanian and Moldavian. I do not offer any explanation.

Paragraph: {}
Dialect:

RoSum

I read the following paragraph and summarize it.

Title: {}
Paragraph: {}
Summary:

SaRoCo

I read and annotate the following title and paragraph based on the satire category.

The only satire categories I can choose from are: satiric and non-satiric. I do not offer any explanation.

Title: {}
Paragraph: {}
Category:

D.2 Original Prompts in Romanian

RoMedQA

Eu răspund la întrebări de medicină de tip grilă doar cu cifra răspunsului corect din variantele de răspuns. Există un singur răspuns corect.

Întrebare: {}
Răspuns:

RoQA

Eu citesc contextul dat și răspund succint la întrebările adresate, utilizând doar informații din context. Nu ofer nicio explicație.

Context: {}
Întrebare: {}
Răspuns:

REDv2

Eu citesc următorul text și îl adnotez în funcție de emoția lui predominanta.

Singurele categorii de emoții din care pot sa aleg sunt: Tristețe, Surpriză, Frică, Furie, Neutru, Încredere, Bucurie. Nu ofer nicio explicație.

Text: {}
Emoție:

RoSTS

Eu citesc ambele propoziții și adnotez similaritatea semantică dintre cele două propoziții cu un scor de la 0 (propozițiile nu au nicio similaritatea semantică) la 1 (propozițiile sunt identice din punct de vedere semantic). Nu ofer nicio explicație.

Propoziție1: {}
Propoziție2: {}
Scor similaritate semantică:

RoMD

Eu citesc următorul paragraf și îl adnotez în funcție de dialectul lui, românesc sau moldovenesc.

Singurele categorii de dialect din care pot sa aleg sunt românesc sau moldovenesc. Nu ofer nicio explicație.

Paragraf: {}
Dialect:

RoSum

Eu citesc următorul paragraf și îl sumarizez.

Titlu: {}
Paragraf: {}
Sumarizare:

SaRoCo

Eu citesc următorul titlu și paragraf, și le adnotez în funcție de categoria de satira.

Singurele categorii de satiră din care pot sa aleg sunt: satiric sau non-satiric. Nu ofer nicio explicație.

Titlu: {}
Paragraf: {}
Categorie: