# "Vorbești Românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions

**Mihai Masala**[a,b,c], **Denis C. Ilie-Ablachim**[b], **Alexandru Dima**[a,b], **Dragos Corlatescu**[b], **Miruna Zavelca**[a,d], **Ovio Olaru**[f], **Simina Terian**[f], **Andrei Terian**[f], **Marius Leordeanu**[b,c], **Horia Velicu**[e], **Marius Popescu**[d], **Mihai Dascalu**[b], **Traian Rebedea**[b,g]
[a]Institute for Logic and Data Science, Bucharest,
[b]National University of Science and Technology POLITEHNICA Bucharest, Romania,
[c]Simion Stoilow Institute of Mathematics of The Romanian Academy, Bucharest,
[d]University of Bucharest, Romania, [e]BRD Groupe Societe Generale,
[f]Lucian Blaga University Sibiu, [g]NVIDIA

## Abstract

In recent years, Large Language Models (LLMs) have achieved almost human-like performance on various tasks. While some LLMs have been trained on multilingual data, most of the training data is in English; hence, their performance in English greatly exceeds other languages. To our knowledge, we are the first to collect and translate a large collection of texts, instructions, and benchmarks and train, evaluate, and release open-source LLMs tailored for Romanian. We evaluate our methods on four different categories, including academic benchmarks, MT-Bench (manually translated), and a professionally built historical, cultural, and social benchmark adapted to Romanian. We argue for the usefulness and high performance of RoLLMs by obtaining state-of-the-art results across the board. We publicly release all resources (i.e., data, training and evaluation code, models) to support and encourage research on Romanian LLMs while concurrently creating a generalizable recipe, adequate for other low or less-resourced languages.

## 1 Introduction

The Transformer architecture (Vaswani et al., 2017) has become ubiquitous recently in the Natural Language Processing (NLP) domain, leading to state-of-the-art performance on various tasks, ranging from text classification to text generation. Decoder-only language models such as GPT series (Radford et al., 2018, 2019; Brown et al., 2020) exhibit impressive capabilities such as understanding and generating natural language, learning and solving tasks not directly trained on. In the last few years, Large Language Models (LLMs) development has exploded, with a plethora of models available (Achiam et al., 2023; Chowdhery et al., 2023; Touvron et al., 2023b; Hoffmann et al., 2022; Taori et al., 2023; Jiang et al., 2023).

Training such massive models requires a vast amount of available data (ranging from tens to hundreds of billions of text tokens) besides computational resources. Intrinsically, most work on such models focuses mainly on English, often excluding other languages. Here, we focus on building the resources for training LLMs specialized in Romanian, a less-resourced language. By showcasing our approach and releasing our models, we strive to enable further research and exciting applications for Romanian-speaking users.

Our contributions can be summarized as follows: **I.** We present and release the first Romanian LLMs with both foundational and instruct variants, based on popular open-source models such as Llama2 (Touvron et al., 2023b), Llama3[1], Mistral (Jiang et al., 2023) and Gemma (Team et al., 2024). Our collection of models is grouped under the OpenLLM-Ro initiative[2]; **II.** We collect and translate a large set of evaluation benchmarks (including a human translation of MT-Bench (Zheng et al., 2024)), conversation, and instructions; **III.** We extend and adapt existing testing suites to thoroughly assess the performance of different models on Romanian tasks. **IV.** We propose a novel methodology to assess language-specific LLMs by proposing the first benchmark that evaluates the Romanian cultural knowledge of LLMs; **V.** We make our recipe (i.e., datasets and code) and models open-source to encourage further research both for Romanian and other low/less-resourced languages.

## 2 Related Work

With the advent of open-source LLMs, especially Llama2 (Touvron et al., 2023b) and Mistral (Jiang et al., 2023), non-English specialized alternatives have been trained. Luukkonen et al. (2023) have

[1]ai.meta.com/blog/meta-llama-3/
[2]www.openllm.ro



*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647
November 12-16, 2024 

trained a family of Finnish foundational models, ranging from 186M to 176B parameters. Trained on a total of 38B tokens with a custom tokenizer, their models showcase state-of-the-art performance on a variety of Finish downstream tasks. Similarly, Cui et al. (2023) have trained both a foundational and a chat variant of Llama2 for Chinese, while Wang (2023) worked on a domain-specific (i.e., media) Chinese LLM. For Italian, Bacciu et al. (2024) translated four out of the six benchmarks used in Open LLM Leaderboard[3] to release the first Italian LLM benchmark and evaluate existing Italian models (Basile et al., 2023; Bacciu et al., 2023; Santilli and Rodolà, 2023; Bacciu et al., 2024), models that are all adapter-based fine-tunes. Other language-specific LLMs include versions for Dutch[4], Bulgarian[5], Serbian (Aleksa, 2024), or Japanese (Fujii et al., 2024).

Dac Lai et al. (2023) automatically translated English Alpaca (Taori et al., 2023) instructions together with a set of generated Self-Instruct (Wang et al., 2023) instructions into 26 languages. Furthermore, they translated for evaluation three out of the six benchmarks used in the Open LLM Leaderboard. They fine-tuned Llama (Touvron et al., 2023a), thus obtaining language-specific models. Li et al. (2023b) translated the Alpaca and Dolly (Conover et al., 2023) datasets into 51 languages and fine-tuned the Llama model, further evaluated in a zero-shot setup on downstream tasks.

Singh et al. (2024) contributed with a huge amount of multilingual text resources under the Aya initiative. Over 2000 people across 119 countries contributed to the creation of Aya Dataset, a dataset of over 204k human-curated instructions in 65 languages. The Aya Collection uses instruction-style templates created by fluent speakers and applies them to existing datasets, generating over 500M instances that cover 114 languages. Based on Aya resources, Üstün et al. (2024) fine-tuned Aya-101, a 13B parameter mT5 model (Xue et al., 2021) capable of following instructions in 101 languages. Compared to Aya-101, which focuses on breadth (i.e., a lot of languages), Aryabumi et al. (2024) considered a more depth-based approach, building Aya-23, a 23-language multilingual instruct model following Cohere's Command model[6].

Based on previous research (Yong et al., 2023; Zhao et al., 2024), we decided to perform both continual pre-training and instruction fine-tuning.

## 3 Data

We employ two distinct data sources for training a Romanian LLM: documents for training the foundational model and instructions for the chat variant.

### 3.1 CulturaX

CulturaX (Nguyen et al., 2023) is a collection of documents in 167 languages, built on top of mC4 (Xue et al., 2021) and OSCAR (Abadji et al., 2022) and using multiple cleaning and deduplication stages to ensure a high-quality corpus. Around 40M of these documents are in Romanian amounting to 40B tokens.

### 3.2 Instructions and Conversations

We collect a large and diverse dataset of conversations and instructions to build the instruct models. Since high-quality data in Romanian is very scarce, we resort to translating[7] datasets such as Alpaca (Taori et al., 2023), Dolly (Conover et al., 2023), NoRobots (Rajani et al., 2023), GPT-Instruct[8], Orca (Mukherjee et al., 2023), Camel (Li et al., 2023a), and UltraChat (Ding et al., 2023). Furthermore, we use Romanian data from OpenAssistant (Köpf et al., 2023) and Aya Collection (Singh et al., 2024). This led to a total of around 2.7M instructions and conversations in Romanian, which were further used in our experiments.

## 4 Training

Initial pre-training experiments were limited to a maximum sequence length of 512 to accelerate training. We further experimented with longer sequence lengths (i.e., 1024 and 2048), always picking the biggest batch size that fits into the memory. For all experiments, we used an AdamW optimizer with weight decay set to 0.1 and gradient clipping. For both continual pre-training and instruction fine-tuning, we performed full training (i.e., updating all model parameters).

### 4.1 Foundational Model

Training the foundational model is done iteratively using continual pre-training on the CulturaX

[3] www.huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[4] www.huggingface.co/BramVanroy/GEITje-7B-ultra
[5] www.huggingface.co/INSAIT-Institute
[6] www.cohere.com/command
[7] www.systransoft.com/ last accessed 6th of March 2024
[8] www.github.com/teknium1/GPTeacher

dataset. Starting from the foundational Llama2 model, we trained the first model on 5% of the CulturaX, the following model on the next 5% (leading to the second model seeing 10% of the data), and the third model on another 10% of the CulturaX. Therefore, we have three variants of a Romanian foundational model, variants that used 5%, 10%, and 20% of the entire CulturaX dataset. Training is done over one epoch, with a fixed learning rate of $1e^{-4}$.

### 4.2 Instruct Models

RoLlama2-Instruct is built by fine-tuning the Romanian Llama2 base model, on one epoch with a fixed learning rate of $5 \times 10^{-5}$. We further built Romanian instruct variants of other popular models (i.e., Mistral-7b, Gemma-7b, and Llama3-8b), dropping the pre-training and directly performing instruction fine-tuning on the base models. For these models, a learning rate of $10^{-6}$ for two epochs worked best.

## 5 Evaluation

Evaluating a general LLM is not straightforward, considering the vast number of benchmarks and evaluation protocols proposed and evaluating LLMs in any language besides English only adds another layer of complexity. We propose an extensive set of evaluations, grouped into four different categories, to obtain a thorough performance assessment of LLMs in Romanian.

### 5.1 Academic Benchmarks

Popular benchmarks include Open LLM Leaderboard, an online leaderboard that encompasses six reasoning and general knowledge tasks: ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), TruthfulQA (Lin et al., 2021), Winogrande (Sakaguchi et al., 2021) and GSM8k (Cobbe et al., 2021).

For Romanian versions of ARC, HellaSwag, MMLU, and TruthfulQA, we resort to ChatGPT translations (Dac Lai et al., 2023), while we translate[9] Winogrande and GSM8k. The generally used few-shot values for ARC, HellaSwag, and MMLU (i.e., 25 shots for ARC, 10 shots for HellaSwag, and 5 shots for MMLU) lead to the prompt having a length equal to or even surpassing 2048 tokens. As some of our models have been trained only on sequences of length 512, 1024, or 2048, we opted to use multiple few-shot settings. More specifically, for ARC, we tested with 0, 1, 3, 5, 10, and 25 shots, MMLU and Winogrande with 0, 1, 3, and 5, and HellaSwag with 0, 1, 3, 5, and 10 shots. For GSM8k, we employ 1, 3, and 5 shot evaluation. Results in Table 1 are averaged over all few shot settings for each benchmark.

### 5.2 MT-Bench

MT-Bench proposes a multi-turn instruction set denoting that strong LLMs (e.g., GPT-4) can act as judges and approximate human preferences in a scalable and somewhat explainable way. While most benchmarks in the Open LLM Leaderboard focus more on the intrinsic knowledge of a model, MT-Bench focuses more on the conversation capability and the quality of the generated answers. In our experiments, we use GPT-4o[10] as a judge. We decided to manually translate MT-Bench because it is a small and highly relevant benchmark for instruct models and it represents a crucial clean test set for Romanian models.

### 5.3 Romanian Downstream Tasks

Besides the standard benchmarks that are used for evaluating LLMs, we further evaluate a series of Romanian downstream tasks to assess the "real-world" performance of our models in solving Romanian tasks. We selected four tasks from the LiRo benchmark (Dumitrescu et al., 2021): sentiment analysis on LaRoSeDa (Tache et al., 2021), machine translation on WMT-16-ro-en dataset (Bojar et al., 2016), question answering on XQuAD-ro (Artetxe et al., 2019) and semantic text similarity on RO-STS (Dumitrescu et al., 2021). We framed sentiment analysis as a classification task, whereas machine translation, question answering, and semantic text similarity are framed as text generation problems. For semantic text similarity, we follow the approach of Gatto et al. (2023), with further details presented in Appendix Section C.

For all tasks, we perform zero/few-shot evaluation (0, 1, 3, and 5 shot averaged) and also fine-tune the models. We used LoRA (Hu et al., 2021) to ensure equitable comparisons between models, targeting all linear layers, with a rank and alpha of 8, a 0.1 dropout, and a learning rate of $5 \times 10^{-5}$. In our initial experiments, we found this approach to generate the most consistent results. We fine-tuned the models for each dataset for only one epoch as

[9] same system used for translating instructions

[10] www.openai.com/index/hello-gpt-4o/ last accessed on $11^{th}$ of June 2024

| Model | Average | ARC | MMLU | Wino | HS | GSM8k | TQA |
|---|---|---|---|---|---|---|---|
| | | | ***Llama2*** | | | | |
| Llama-2-7b | 37.04 | 36.05 | 33.66 | 57.56 | 48.00 | 4.75 | 42.22 |
| *RoLlama2-7b-Base* (512-5%) | 38.13 | 37.66 | 31.56 | 59.51 | 56.25 | 3.29 | 40.51 |
| *RoLlama2-7b-Base* (1024-5%) | 39.20 | 37.99 | 31.99 | 59.71 | 56.91 | 4.12 | 44.51 |
| *RoLlama2-7b-Base* (2048-5%) | **39.31** | 38.35 | 30.29 | 59.30 | 57.36 | 5.08 | 45.48 |
| *RoLlama2-7b-Base* (1024-10%) | 38.42 | 38.09 | 30.28 | 59.00 | 57.58 | 2.98 | 42.58 |
| *RoLlama2-7b-Base* (1024-20%) | 38.03 | 37.95 | 27.22 | 59.29 | 57.22 | 2.53 | 44.00 |
| Llama-2-7b-chat | 36.84 | 37.03 | 33.81 | 55.87 | 45.36 | 4.90 | 44.09 |
| *RoLlama2-7b-Instruct* | **44.50** | 44.73 | 40.39 | 63.67 | 59.12 | 13.29 | 45.78 |
| | | | ***Mistral*** | | | | |
| Mistral-7B-v0.1 | 45.02 | 42.99 | 47.16 | 60.77 | 54.19 | 16.20 | 48.80 |
| Mistral-7B-Instruct-v0.2 | 47.40 | 46.29 | 47.01 | 58.78 | 54.27 | 13.47 | **64.59** |
| *RoMistral-7b-Instruct* | **52.91** | **52.27** | 49.33 | **70.03** | **62.88** | 32.42 | 50.51 |
| | | | ***Llama3*** | | | | |
| Llama-3-8B | 44.55 | 38.05 | 48.33 | 59.94 | 53.48 | 20.04 | 47.44 |
| Llama-3-8B-Instruct | 50.62 | 43.69 | 52.04 | 59.33 | 53.19 | 43.87 | 51.59 |
| *RoLlama3-8b-Instruct* | **52.21** | 47.95 | 53.50 | 66.06 | 59.72 | 40.16 | 45.90 |
| | | | ***Llama3.1*** | | | | |
| Llama-3.1-8B | 37.29 | 33.25 | 36.35 | 58.80 | 42.65 | 3.59 | 49.03 |
| Llama-3.1-8B-Instruct | 49.87 | 42.86 | 53.73 | 59.71 | 56.82 | 35.56 | 50.54 |
| *RoLlama3.1-8b-Instruct* | **53.03** | 47.69 | **54.57** | 65.85 | 59.94 | **44.30** | 45.82 |
| | | | ***Gemma*** | | | | |
| gemma-7b | 50.04 | 47.22 | 53.18 | 61.46 | 60.32 | 30.48 | 47.59 |
| gemma-1.1-7b-it | 41.39 | 40.05 | 47.12 | 54.62 | 47.10 | 9.73 | 49.75 |
| *RoGemma-7b-Instruct* | **50.48** | 52.02 | 52.37 | 66.97 | 56.34 | 25.98 | 49.18 |
| | | | ***Other models*** | | | | |
| Okapi-Ro | 35.64 | 37.90 | 27.29 | 55.51 | 48.19 | 0.83 | 44.15 |
| aya-23-8B | **45.81** | 43.89 | 45.96 | 60.50 | 60.52 | 16.81 | 47.16 |

Table 1: Comparison between RoLLMs and other LLMs on Romanian versions of academic benchmarks (abbreviations: (1024-10) - sequence length of 1024, 10% of CulturaX used for pretraining, HS - HellaSwag, Wino - Winogrande, TQA - TruthfulQA). **Bold** denotes the best in each category (average) and overall (each benchmark).

we found that this was more than enough for the loss to stabilize.

### 5.4 RoCulturaBench

Finally, we were interested in how grounded are LLMs in the historical, cultural, and social realities of Romania. For this reason, we devised a new benchmark, in the form of RoCulturaBench. RoCulturaBench was built manually by a team of Romanian academics from the field of humanities and aims to address all the significant aspects of Romanian culture (in the broad sense of the term, ranging from artistic and scientific contributions to cuisine and sports). The benchmark consists of 100 prompts grouped into 10 domains, covering three types of data (see Appendix Section B for examples): (a) communicating information in various formal and informal contexts (the first three categories, concerning grammar and spell check, expressing subjectivity – especially emotions and personal values –, producing functional texts, role-playing and impersonating another person, creativity, and different ways of evaluation and argumentation); (b) factual information about Romania (categories IV to VII, which refer to Romanian geography, tourism, history, demography, politics, science, culture, and sports); (c) information on how Romanians perceive themselves and the world (e.g., traditions, customs, beliefs, stereotypes, representations, attitudes, values).

The design of the benchmark has intentionally avoided any obvious and/or unequivocal answers, such as those resulting from the usage of scientific laws or formulas. The instructions included notions and qualifiers that would each time prompt human subjects to consider at least two possible

| Model | Average | 1st turn | 2nd turn | #Answers in Ro |
|---|---|---|---|---|
| Llama-2-7b-chat | 0.59 | 0.95 | 0.23 | 21 / 160 |
| Llama-2-7b-chat (*ro_prompted*) | 1.08 | 1.44 | 0.73 | 45 / 160 |
| *RoLlama2-7b-Instruct* | **4.43** | 4.93 | 3.94 | **160 / 160** |
| Mistral-7B-Instruct-v0.2 | 1.04 | 1.35 | 0.72 | 27 / 160 |
| Mistral-7B-Instruct-v0.2 (*ro_prompted*) | 5.03 | 5.05 | 5.00 | 154 / 160 |
| *RoMistral-7b-Instruct* | **5.29** | 5.86 | 4.73 | **160 / 160** |
| Llama-3-8B-Instruct | 4.54 | 4.65 | 4.43 | 126 / 160 |
| Llama-3-8B-Instruct (*ro_prompted*) | **5.96** | **6.16** | **5.76** | 158 / 160 |
| *RoLlama3-8b-Instruct* | 5.38 | 6.09 | 4.68 | **160 / 160** |
| Llama-3.1-8B-Instruct | **5.69** | 5.96 | 5.43 | 160 / 160 |
| Llama-3.1-8B-Instruct (*ro_prompted*) | **5.69** | 5.85 | 5.53 | 160 / 160 |
| *RoLlama3.1-8b-Instruct* | 5.42 | 5.95 | 4.89 | **160 / 160** |
| gemma-1.1-7b-it | 5.09 | 5.53 | 4.66 | 158 / 160 |
| gemma-1.1-7b-it (*ro_prompted*) | 4.83 | 5.11 | 4.55 | **160 / 160** |
| *RoGemma-7b-Instruct* | **5.24** | 5.55 | 4.94 | **160 / 160** |

Table 2: Romanian MT-Bench performance. **Bold** as in Table 1.

answers for each prompt (depending on how they understood the notions and qualifiers involved in the prompt) but, at the same time, limit the instructions' field of reference so as to not allow a broad spectrum of possible answers.

We provide reference answers for all prompts and evaluate the LLM-generated answers using GPT-4o[11] as a judge. Compared to MT-Bench, RoCulturaBench contains only single-turn interactions (i.e., one instruction, one answer).

## 6 Results & Discussions

In this section, we present and discuss the results of our models on the four evaluations: academic benchmarks, MT-Bench, Romanian downstream tasks, and RoCulturaBench.

### 6.1 Academic Benchmarks

Table 1 introduces the academic benchmark results. In the first section, Llama2, and different versions of RoLlama2 foundational model are compared. Increasing the pretraining sequence length boosts the overall performance while adding more pretraining data (up to 20% of our data) slighty degrades the foundational model's performance. We conjecture that this is due to the quality of the pretraining data used, as CulturaX contains data from different sources, including low-quality sources. Nevertheless, we found foundational models exposed to more training data responding better to instruction fine-tuning, consistently outperforming models trained on less data. For that reason, even if the model has lower performance than some of its counterparts, we perform instruction fine-tuning on RoLlama2-7b-Base that was trained on 20% of the data with a sequence length of 1024. Still, pre-training on Romanian texts increases the performance on four of the six benchmarks when compared to the Llama2 foundational model. There is a noticeable increase in the Winogrande and HellaSwag benchmarks, arguing for the model's capability to understand Romanian and use it for reasoning. On GSM8k and MMLU, we observe a consistent drop in performance that could be explained by the lack of appropriate data in pre-training, thus making it easier for the model to forget mathematical reasoning (for GSM8k) and knowledge about US history or law. Fine-tuning with instructions further increases the overall score by around 6% over RoLlama2-base with an increase on all of the six benchmarks.

Turning to another family of models, we notice a significant increase in the performance of the Mistral, Llama3, and Gemma models in Romanian. While we did not find details regarding the languages of the training data for Mistral, Gemma and Llama3 were trained mainly on English texts; as such, their proficiency in understanding Romanian is somewhat surprising. Nevertheless, further exposure to Romanian instructions boosts performance across the board, from 3% up to 8%.

Finally, we evaluate other existing models for Romanian including Okapi-Ro, an LLM specialized in Romanian, and Aya, a multilingual LLM

[11] gpt-4o-2024-08-06

| Model | Zero/Few shot | | Fine-tuned | |
|---|---|---|---|---|
| | Binary (*F1*) | Multiclass (*F1*) | Binary (*F1*) | Multiclass (*F1*) |
| SOTA - Romanian BERT[a] | - | - | 98.31 | 86.63 |
| Llama-2-7b | 93.19 | 54.11 | 98.43 | 87.22 |
| *RoLlama2-7b-Base* | 83.25 | 61.04 | 98.97 | 87.72 |
| Llama-2-7b-chat | 87.78 | 52.82 | 97.27 | 82.02 |
| *RoLlama2-7b-Instruct* | **97.66** | 62.41 | 97.97 | 60.89 |
| Mistral-7B-Instruct-v0.2 | 96.97 | 56.66 | 98.83 | 87.32 |
| *RoMistral-7b-Instruct* | 95.56 | **67.83** | **99.00** | **87.57** |
| Llama-3-8B-Instruct | 95.88 | 56.21 | 98.53 | 86.19 |
| *RoLlama3-8b-Instruct* | 95.58 | 61.20 | 96.46 | 87.26 |
| Llama-3.1-8B-Instruct | 95.74 | 59.49 | 98.57 | 82.41 |
| *RoLlama3.1-8b-Instruct* | 94.57 | 60.10 | 95.12 | 87.53 |
| gemma-1.1-7b-it | 87.54 | 51.49 | 83.87 | 85.61 |
| *RoGemma-7b-Instruct* | 86.96 | 56.72 | 98.80 | 85.81 |
| GPT-3.5* | 83.64 | 44.70 | - | - |
| GPT-4o* | 92.78 | 45.22 | - | - |

Table 3: LaRoSeDa (sentiment analysis) performance. **Bold** denotes the best result for each setting. *Unlike other models where we framed the problem as a classification task (i.e., comparing which class has the highest logprob), we treat the problem as a generation task for GPT, filtering the generated text. [a]Niculescu et al. (2021)

with support for Romanian. As Okapi is a model based on the first Llama model, its lackluster performance is expected. Aya-23 has official support for 23 languages, including Romanian, and it performs on par with the instruct version of RoLlama2 and with Mistral models.

### 6.2 MT-Bench

Performance on MT-Bench is presented in Table 2. We noticed that even when using a Romanian prompt, most Llama2, Mistral, and Llama3 answers, while somewhat correct, are either entirely in English or have the first few words in Romanian, then reverting to English. We specifically ask the judge to penalize answers that are not in Romanian, but unfortunately, experiments showed that the GPT judge scarcely does this, assigning the same score to English answers irrespective of our request. For this reason, we decided to automatically detect the language of the answers[12] and penalize the answers that are not written in Romanian with a score of 0. Upon manual inspection, we decided to manually assign the language using simple rules for a few specific prompts where automated methods fail. To alleviate the predisposition of models to answer in English even if the prompt is in Romanian, we added after each instruction an additional request to answer only in Romanian. As seen in the last column in Table 2, when specifically prompted to answer in Romanian (*ro_prompted*), the models tended to perform better, with Llama3 and Gemma being capable of answering mostly in the correct language. As expected, this is not an issue with RoLLMs, as they generate text in Romanian without being specifically told to do so.

Turning to the evaluation scores, we notice the gap in performance between the 1st and 2nd turns is larger for RoLLMs than for their counterparts. For RoLLMs, this is somewhat expected as the instruction fine-tuning dataset contains, for the most part, single-turn conversations. Adding more multi-turn conversations in the fine-tuning dataset should also close the performance gap on the second turn, thus leading to greater overall performance.

We also emphasize that we performed instruction fine-tuning on base models except for the newer Llama3 and Llama3.1 models, for which we lost valuable time and resources spent on alignment. For the Llama3 model family, we found better results by performing direct instruction fine-tuning on the Instruct variant, a model fine-tuned and aligned with human preferences on a huge dataset that also contains 10M human-annotated examples[13]. To put things into perspective, we translated and used around 2.7M English instructions, out of which only 25k were human-created. As such, we are still investigating how to best train

[12]www.pypi.org/project/langdetect/

[13]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, last accessed on 1st of October 2024

| Model | Zero/Few shot | | Fine-tuned | |
|---|---|---|---|---|
| | EN->RO (*Bleu*) | RO->EN (*Bleu*) | EN->RO (*Bleu*) | RO->EN (*Bleu*) |
| SOTA - mBART[a] | - | - | **38.50** | 39.90 |
| Llama-2-7b | 14.90 | 26.61 | 24.95 | 39.09 |
| *RoLlama2-7b-Base* | 10.01 | 13.03 | 27.85 | 39.30 |
| Llama-2-7b-chat | 15.55 | 28.53 | 19.99 | 31.48 |
| *RoLlama2-7b-Instruct* | 27.14 | 19.40 | 27.63 | 39.75 |
| Mistral-7B-Instruct-v0.2 | 18.60 | 33.99 | 26.19 | 39.88 |
| *RoMistral-7b-Instruct* | 28.28 | 6.10 | 27.70 | 40.36 |
| Llama-3-8B-Instruct | 18.89 | 30.98 | 28.02 | 40.28 |
| *RoLlama3-8b-Instruct* | 22.92 | 24.28 | 27.31 | **40.52** |
| Llama-3.1-8B-Instruct | 19.01 | 27.77 | 29.02 | 39.80 |
| *RoLlama3.1-8b-Instruct* | 21.88 | 23.99 | 28.27 | 40.44 |
| gemma-1.1-7b-it | 17.96 | 27.74 | 25.48 | 36.11 |
| *RoGemma-7b-Instruct* | 24.45 | 14.20 | 25.96 | 39.07 |
| GPT-3.5 | 31.83 | **40.61** | - | - |
| GPT-4o | **34.49** | 39.41 | - | - |

Table 4: WMT'16 (machine translation) performance. **Bold** as in Table 3. [a]Liu et al. (2020).

such a model, including continual instruction fine-tuning (Chen and Lee, 2024).

### 6.3 Romanian Downstream Tasks

Downstream tasks performance is presented in Tables 3, 4, 5 and 6. For sentiment analysis (Table 3), we note the strong performance of RoLLMs compared to their counterparts in both settings. The gap in performance shrinks for fine-tuned variants, as all models are somewhat close in performance. It was previously shown that smaller BERT-based models (Avram et al., 2022; Masala et al., 2020) also achieve comparable performance when fine-tuned (Niculescu et al., 2021), especially for the binary case, suggesting that this might be the limit on this particular dataset.

In the case of machine translation (see Table 4), we note the discrepancy in translation performance between EN->RO and RO->EN in zero/few-shot setting: RoLLMs perform better in EN->RO setting, while their counterparts are more proficient in RO to EN translations. When finetuned, RoLLMs consistently outperform their counterparts, a trend also observed in the sentiment analysis task. The state-of-the-art model for EN<->RO translations is mBART (Liu et al., 2020), a sequence-to-sequence denoising auto-encoder that is pre-trained on large-scale English and Romanian dataset and further fine-tuned on WMT.

In the case of question answering (see Table 5), the fine-tuning setup is different due to the nature of the training data as the XQuAD dataset consists of 1190 question-answer pairs professionally translated from the development set of SQuAD v1.1 (Rajpurkar et al., 2016). Therefore, fine-tuning in this case is done on the English dataset, while evaluation is done on the Romanian question-answer pairs. Naturally, non-RO LLMs tend to respond better to data in English. Also, we note that the foundational models perform better than instruct variants in this particular case of XQuAD. In zero/few-shot settings, RoLLMs perform better, with the exception of RoGemma-7b-Instruct and RoLlama3-8b-Instruct, both models exhibiting very strong zero-shot performance and a downward trend especially with 3 and 5 examples.

Finally, we notice the same pattern in the case of semantic text similarity (see Table 6): strong performance of RoLLMs both when fine-tuned and in a few-shot context. We found in our experiments that most models struggle in the zero-shot setting in this task, generating negative correlation scores. For this reason, we decided to evaluate the semantic text similarity task only on 1, 3, and 5-shot settings, akin to how we evaluated GSM8k. For tasks where the output is numerical and has to adhere to a fixed format, we recommend using at least 1 positive example when using these models.

### 6.4 RoCulturaBench

Results on RoCulturaBench are presented in Table 7. Taking an English foundational model and adapting it to a new language by performing instruction fine-tuning is not enough to properly un-

| Model | Zero/Few shot | | Fine-tuned | |
|---|---|---|---|---|
| | (*Exact Match*) | (*F1*) | (*Exact Match*) | (*F1*) |
| SOTA - XLM-R[a] | - | - | **69.70** | 83.60 |
| Llama-2-7b | 38.91 | 56.82 | 65.46 | 79.42 |
| *RoLlama2-7b-Base* | 30.15 | 47.03 | 67.06 | 79.96 |
| Llama-2-7b-chat | 32.35 | 54.00 | 60.34 | 75.98 |
| *RoLlama2-7b-Instruct* | **45.71** | **65.08** | 59.24 | 74.25 |
| Mistral-7B-Instruct-v0.2 | 27.92 | 50.71 | 65.46 | 79.73 |
| *RoMistral-7b-Instruct* | 41.09 | 63.21 | 47.56 | 62.69 |
| Llama-3-8B-Instruct | 39.47 | 58.67 | 67.65 | 82.77 |
| *RoLlama3-8b-Instruct* | 18.89 | 31.79 | 50.84 | 65.18 |
| Llama-3.1-8B-Instruct | 44.96 | 64.45 | 69.50 | **84.31** |
| *RoLlama3.1-8b-Instruct* | 13.59 | 23.56 | 49.41 | 62.93 |
| gemma-1.1-7b-it | 42.10 | 62.30 | 60.34 | 77.40 |
| *RoGemma-7b-Instruct* | 26.03 | 41.58 | 46.72 | 60.79 |
| GPT-3.5 | 35.53 | 60.68 | - | - |
| GPT-4o | 19.31 | 42.89 | - | - |

Table 5: XQuAD (question answering) performance. **Bold** as in Table 3. [a]Niculescu et al. (2021)

| Model | Few shot | | Fine-tuned | |
|---|---|---|---|---|
| | (*Spearman*) | (*Pearson*) | (*Spearman*) | (*Pearson*) |
| SOTA - Romanian BERT[a] | - | - | 83.15 | 83.70 |
| Llama-2-7b | 9.08 | 9.07 | 79.93 | 81.08 |
| *RoLlama2-7b-Base* | 7.89 | 7.98 | 71.75 | 71.99 |
| Llama-2-7b-chat | 32.56 | 31.99 | 74.08 | 72.64 |
| *RoLlama2-7b-Instruct* | 59.69 | 57.16 | 84.66 | 85.07 |
| Mistral-7B-Instruct-v0.2 | 62.62 | 60.86 | 84.92 | 85.44 |
| *RoMistral-7b-Instruct* | 78.47 | 77.24 | 87.28 | 87.88 |
| Llama-3-8B-Instruct | 73.04 | 72.36 | 83.49 | 84.06 |
| *RoLlama3-8b-Instruct* | 77.60 | 76.86 | 86.70 | 87.09 |
| Llama-3.1-8B-Instruct | 72.11 | 71.64 | 84.59 | 84.96 |
| *RoLlama3.1-8b-Instruct* | 75.89 | 76.00 | 86.86 | 87.05 |
| gemma-1.1-7b-it | 49.10 | 50.23 | 83.43 | 83.64 |
| *RoGemma-7b-Instruct* | 73.23 | 71.58 | **88.42** | **88.45** |
| GPT-3.5 | 79.08 | 78.97 | - | - |
| GPT-4o | **84.36** | **84.36** | - | - |

Table 6: Ro-STS (semantic text similarity) performance. **Bold** as in Table 3. [a] (Niculescu et al., 2021)

.

derstand the nuances and realities of a culture. Testament stands the competitive performance of RoLlama2: starting from an arguably weaker model but also pre-training it performs on par with newer and stronger models that have only been fine-tuned.

We argue that, as expected, cultural information and social knowledge are primarily injected into the model in the pre-training phase, not in the fine-tuning phase, especially when fine-tuning exploits non-natively data (i.e., written in English and then translated). As in the case of MT-Bench, Llama3 outperforms other models by a noteworthy margin, with only GPT-3.5 being stronger. For more detailed results and comparisons, see Appendix Section A.2. In addition, results obtained using the LLM-as-a-judge approach are not the be-all and end-all of evaluations, especially regarding non-English language and culture. We leave human evaluation and other approaches (e.g., training a language-specific judge) for future work.

### 6.5 Human Alignment

After manually inspecting part of the results for RoMTBench and RoCulturaBench, we found that

| Model | Score | #Answers in Ro |
|---|---|---|
| Llama-2-7b-chat (*ro_prompted*) | 1.21 | 33 / 100 |
| *RoLlama2-7b-Instruct* | 4.08 | **100 / 100** |
| Mistral-7B-Instruct-v0.2 (*ro_prompted*) | 3.68 | 97 / 100 |
| *RoMistral-7b-Instruct* | 3.99 | **100 / 100** |
| Llama-3-8B-Instruct (*ro_prompted*) | **4.62** | **100 / 100** |
| *RoLlama3-8b-Instruct* | 3.81 | **100 / 100** |
| Llama-3.1-8B-Instruct (*ro_prompted*) | 3.54 | **100 / 100** |
| *RoLlama3.1-8b-Instruct* | **3.55** | **100 / 100** |
| gemma-1.1-7b-it (*ro_prompted*) | 3.38 | **100 / 100** |
| *RoGemma-7b-Instruct* | **3.51** | **100 / 100** |
| GPT-3.5 | **6.09** | **100 / 100** |

Table 7: RoCulturaBench performance. **Bold** denotes as in Table 1.

| Model | RoMT-Bench | RoCulturaBench |
|---|---|---|
| RoLlama2-7b-Instruct | 4.43 | 4.08 |
| RoLlama2-7b-Instruct-DPO | 4.61 | **4.80** |
| RoMistral-7b-Instruct | 5.29 | 3.99 |
| RoMistral-7b-Instruct-DPO | 5.88 | 4.72 |
| RoLlama3-8b-Instruct | 5.38 | 3.81 |
| RoLlama3-8b-Instruct-DPO | 5.87 | 4.40 |
| RoLlama3.1-8b-Instruct | 5.42 | 3.55 |
| RoLlama3.1-8b-Instruct-DPO | **6.21** | 4.42 |
| RoGemma-7b-Instruct | 5.24 | 3.51 |
| RoGemma-7b-Instruct-DPO | 5.48 | 3.94 |

Table 8: Performance of aligned RoLLMs. **Bold** denotes the best results for each benchmark.

our models tend to have a more rigid answer. The other models have a more conversational tone, which is more in line with the reference answers and general human preferences. This sometimes rigid tone of RoLLMs stems mostly from the instruction dataset used for fine-tuning and the lack of human preference alignment mechanisms. To test this, we decided to translate the HelpSteer dataset (Wang et al., 2024) containing human preference data and use it to perform Direct Preference Optimization (Rafailov et al., 2024) (DPO) on our models.

Results are presented in Table 8. Using only 15k samples, we notice a significant increase in performance across the board, proving the essential role of human alignment for LLMs. We leave a more thorough approach for human alignment, including using more data, to future work.

## 7 Conclusions

This paper introduces the first open-source LLM models specialized for Romanian. The released models show promising results, outperforming existing solutions for a wide array of tasks.

We present a general training and evaluation recipe which we expect to work on different, more powerful LLMs as well (e.g., models larger than 7-8B, Mixtral[14] models) as well as for other languages. We firmly believe that this is just the first step towards building and adapting LLMs for Romanian for both research and industry applications.

Pretraining on additional cleaner data, validating existing translations, and adding more instructions, conversations and human preference data are just a few improvements to increase the quality of RoLLMs.

## Limitations

Our approach for adapting existing LLMs to a new language (i.e., Romanian) has some notable limitations. One significant drawback is the reliance on translated (mainly single-turn) instructions for the instruction fine-tuning step. Subtle nuances might not be captured by the translation process, a process that might introduce inaccuracies that de-

[14] www.mistral.ai/news/mixtral-of-experts/

crease the model's performance. Additionally, our models lack any kind of safeguard or moderation mechanism, meaning that they could generate inappropriate, racist, toxic, dangerous, or even illegal content, raising ethical and practical concerns.

## Acknowledgements

Synchronization and collaboration for this work took place within a project[15] carried out in the Institute of Logic and Data Science with funding from BRD Groupe Societe Generale. Part of this work was funded by the EU's NextGenerationEU instrument through the National Recovery and Resilience Plan of Romania – Pillar III-C9-I8, managed by the Ministry of Research, Innovation and Digitalization, within the project entitled Measuring Tragedy: Geographical Diffusion, Comparative Morphology, and Computational Analysis of European Tragic Form (METRA), contract no. 760249/28.12.2023, code CF 163/31.07.2023. We thank Alin Stefanescu and Laurentiu Leustean for participating in various discussions regarding this project.

## References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Gordić Aleksa. 2024. Yugogpt - an open-source llm for serbian, bosnian, and croatian languages.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Andrei-Marius Avram, Darius Catrina, Dumitru-Clementin Cercel, Mihai Dascalu, Traian Rebedea, Vasile Pais, and Dan Tufis. 2022. Distilling the knowledge of Romanian BERTs using multiple teachers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 374–384, Marseille, France. European Language Resources Association.

Andrea Bacciu, Cesare Campagnano, Giovanni Trappolini, and Fabrizio Silvestri. 2024. DanteLLM: Let's push Italian LLM research forward! In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4343–4355, Torino, Italia. ELRA and ICCL.

Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. 2023. Fauno: The italian large language model that will leave you senza parole! *arXiv preprint arXiv:2306.14457*.

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*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kuang-Ming Chen and Hung-yi Lee. 2024. Instructioncp: A fast approach to transfer large language models into target language. *arXiv preprint arXiv:2405.20175*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind

[15] www.ilds.ro/llm-for-romanian

Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems, 2021. *URL https://arxiv.org/abs/2110.14168*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

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

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Joseph Gatto, Omar Sharif, Parker Seegmiller, Philip Bohlman, and Sarah Preum. 2023. Text encoders lack knowledge: Leveraging generative LLMs for domain-specific semantic textual similarity. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 277–288, Singapore. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

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*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large scale language model society. *Preprint*, arXiv:2303.17760.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, et al. 2023. Fingpt: Large generative models for a small language. *arXiv preprint arXiv:2311.05640*.

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.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. 2023. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

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

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Anca Tache, Gaman Mihaela, and Radu Tudor Ionescu. 2021. Clustering word embeddings with self-organizing maps. application on LaRoSeDa - a large Romanian sentiment data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 949–956, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

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.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2024. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. 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 3371–3384, Mexico City, Mexico. Association for Computational Linguistics.

Zhonghao Wang. 2023. Mediagpt: A large language model target chinese media. *arXiv preprint arXiv:2307.10930*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and

Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

# A Detailed results

## A.1 MT-Bench

The scores for each category for Romanian MT-Bench are presented in Table 9. We notice a general upward trend for more scientific categories such as math, extraction (prompts akin to downstream tasks), stem, and with the exception of RoLLama3 even for the writing category. On the other hand, our models exhibit degrading performance in coding, reasoning, and roleplaying categories. For the roleplay and code categories, this is expected due to the nature of our instruction dataset (i.e., mostly single-turn conversations and no coding instructions). We believe that a larger, more diverse set of instructions should easily bridge the performance gap in these categories, and even inverse the downward trend.

## A.2 RoCulturaBench

In Table 10, we present results on RoCulturaBench with and without taking into account the human-written reference answers. First, we notice an overall lower score when the judge is given the reference answer and is asked to compare it to the output of a model. This decrease in performance is consistent and easily noticeable for every model. But this decrease is not completely uniform: for RoLLMs, the penalty tends to be lower (e.g., in the case of Llama3, we notice a loss of 1.53, while for RoLLama3, the score is lower by 1.23).

Scores for each category of RoCulturaBench are presented in Table 11. We note the very strong performance of GPT-3.5 of Llama3 across all categories. The Role category seems to be the closest category in terms of scores, while for the Geography and Sports categories, both GPT-3.5 and Llama3 really shine compared to all other models.

# B RoCulturaBench Example

In Table 12 we present a few examples prompts from each category of RoCulturaBench in both Romanian and English.

# C Prompts for Downstream Tasks

The prompts used for the for downstream tasks are available in Table 13. The models received only the Romanian prompts, however we also included their translations into English for increased readability.

# D Model Architecture

As we continue the process of pre-training or instruction fine-tuning on the foundational model, for each RoLLM we inherit the architecture and the vocabulary of the English variant. Specifically, RoLlama2-7b models use the standard Transformer (Vaswani et al., 2017) architecture with 32 attention heads, 32 layers, hidden layers of size 4096, and a vocabulary of size 32000. Mistral shares, for the most part, the same architecture with Llama2, adding a Sliding Window Attention (Beltagy et al., 2020) mechanism, keeping the same vocabulary. Llama3-8B and, by extension RoLlama3-8b-Instruct, adds grouped-query attention (Ainslie et al., 2023) on top of the Llama2 architecture and also increases the vocabulary size up to 128k. Finally, RoGemma-7b uses 28 layers, 16 attention heads, hidden layers of size 3072 and a vocabulary size of 256k tokens.

# E Computational Resources

Pre-training was performed on up to six NVIDIA A100s with 80GB of memory. Continual pre-training on 5% of the data using a sequence length

| Model | W | Rp | R | M | Code | Ext | STEM | H |
|---|---|---|---|---|---|---|---|---|
| Llama-2-7b-chat (*ro_prompted*) | 1.95 | 1.15 | 0.70 | 1.00 | 0.25 | 1.35 | 1.35 | 0.90 |
| *RoLlama2-7b-Instruct-DPO* | 5.60 | 6.25 | 4.35 | 2.30 | 2.25 | 3.95 | 5.35 | 6.80 |
| Mistral-7B-Instruct-v0.2 (*ro_prompted*) | 4.70 | 5.00 | 5.40 | 3.80 | 4.85 | 3.65 | 5.50 | 7.30 |
| *RoMistral-7b-Instruct-DPO* | 6.75 | 6.50 | 6.90 | 4.50 | 3.35 | **5.60** | 5.85 | **7.60** |
| Llama-3-8B-Instruct (*ro_prompted*) | **7.20** | 6.25 | 7.55 | 4.30 | 5.40 | 4.70 | 5.70 | 6.60 |
| *RoLlama3-8b-Instruct-DPO* | 6.90 | 6.60 | 7.25 | 4.80 | 3.35 | 4.55 | **6.60** | 6.90 |
| Llama-3.1-8B-Instruct (*ro_prompted*) | 6.55 | 4.95 | 7.35 | 5.50 | **5.75** | 3.55 | 5.25 | 6.63 |
| *RoLlama3.1-8b-Instruct-DPO* | 6.85 | **6.65** | **7.90** | **6.25** | 3.95 | 4.55 | 6.00 | 7.55 |
| gemma-1.1-7b-it (*ro_prompted*) | 5.45 | 5.00 | 5.75 | 3.45 | 3.80 | 4.05 | 5.00 | 6.15 |
| *RoGemma-7b-Instruct-DPO* | 5.90 | 5.95 | 6.35 | 5.35 | 3.55 | 4.55 | 5.65 | 6.60 |

Table 9: Performance on each category of Romanian MT-Bench. W - Writing, Rp - Roleplay, R - Reasoning, M - Math, Code - Coding, Ext - Extraction, H - Humanities. **Bold** denotes the best result for each category.

| Model | W/ Ref | W/o Ref | Score diff |
|---|---|---|---|
| Llama-2-7b-chat (*ro_prompted*) | 1.21 | 2.42 | 1.21 |
| *RoLlama2-7b-Instruct-DPO* | 4.80 | 5.95 | 1.15 |
| Mistral-7B-Instruct-v0.2 (*ro_prompted*) | 3.68 | 4.48 | 0.80 |
| *RoMistral-7b-Instruct-DPO* | 4.72 | 5.88 | 1.16 |
| Llama-3-8B-Instruct (*ro_prompted*) | 4.62 | 5.60 | 0.98 |
| *RoLlama3-8b-Instruct-DPO* | 4.40 | 5.38 | 0.98 |
| Llama-3.1-8B-Instruct (*ro_prompted*) | 3.54 | 4.50 | 0.96 |
| *RoLlama3.1-8b-Instruct-DPO* | 4.42 | 5.51 | 1.09 |
| gemma-1.1-7b-it (*ro_prompted*) | 3.38 | 4.07 | 0.69 |
| *RoGemma-7b-Instruct-DPO* | 3.94 | 4.71 | 0.77 |

Table 10: RoCulturaBench performance with and without reference answers.

| Model | W | R | A | G | H | C&S | Sp | C&T | S | C |
|---|---|---|---|---|---|---|---|---|---|---|
| Llama-2-7b-chat (*ro_prompted*) | 0.0 | 1.5 | 1.5 | 1.6 | 2.2 | 1.7 | 0.3 | 0.6 | 1.4 | 1.3 |
| *RoLlama2-7b-Instruct-DPO* | 5.4 | 6.0 | 6.0 | 5.2 | 4.8 | 3.9 | 3.4 | 4.7 | 4.3 | 4.3 |
| Mistral-7B-Instruct-v0.2 (*ro_prompted*) | 4.1 | 3.3 | 5.6 | 3.6 | 3.7 | 3.5 | 3.1 | 2.7 | 4.5 | 2.7 |
| *RoMistral-7b-Instruct-DPO* | 5.6 | 5.7 | 7.1 | 5.5 | 3.6 | 2.9 | 4.3 | 4.1 | 4.9 | 3.5 |
| Llama-3-8B-Instruct (*ro_prompted*) | 5.4 | 5.8 | 6.7 | 4.9 | 3.8 | 3.5 | 3.3 | 4.2 | 5.0 | 3.6 |
| *RoLlama3-8b-Instruct-DPO* | 5.3 | 5.6 | 5.8 | 5.5 | 3.7 | 2.9 | 3.8 | 4.0 | 4.6 | 2.8 |
| Llama-3.1-8B-Instruct (*ro_prompted*) | 4.1 | 4.3 | 5.3 | 3.6 | 2.7 | 2.6 | 3.0 | 3.1 | 4.0 | 2.7 |
| *RoLlama3.1-8b-Instruct-DPO* | 4.9 | 5.7 | 6.8 | 5.2 | 3.5 | 3.4 | 3.1 | 3.7 | 4.7 | 3.2 |
| gemma-1.1-7b-it (*ro_prompted*) | 3.9 | 4.8 | 5.6 | 3.6 | 3.2 | 2.3 | 2.5 | 2.7 | 3.4 | 1.8 |
| *RoGemma-7b-Instruct-DPO* | 5.0 | 5.0 | 6.5 | 3.8 | 3.4 | 2.2 | 3.0 | 3.5 | 4.9 | 2.1 |
| Gpt-3.5 | **6.0** | **7.0** | **7.8** | **7.1** | **5.7** | **4.8** | **4.9** | **6.6** | **5.8** | **5.2** |

Table 11: Performance on each category of RoCulturaBench. W - Writing, R - Roles, A - Argumentation, G - Geography, H - History, C&S - Culture & Science, C&T - Customs & Traditions, Sp - Sports, S - Stereotypes, C - Celebrity. **Bold** denotes the best result for each category.

of 512 for one epoch used 320 GPU hours, 468 GPU hours were consumed for a sequence length of 1024, while for a sequence length of 2048, 708 GPU hours were needed. Instruction fine-tuning of RoLlama2 with a sequence length of 1024 for one epoch required 120 GPU hours, while performing DPO required 30 GPU minutes for one epoch. Evaluations were performed on a single A100 GPU at a time.

| Category | Romanian Prompt | English Prompt |
|---|---|---|
| Writing | Redactează, în calitate de angajat, o cerere în care să-i soliciți într-un mod convingător managerului filialei locale a companiei la care lucrezi să-ți aprobe organizarea unei petreceri cu ocazia zilei de 8 Martie. Precizează că petrecerea se va desfășura în Sala de Conferințe, va începe cu două ore înainte de terminarea programului, va include 50 de participanți și va avea nevoie de un buget de 5.000 de lei din fondul de protocol al filialei. | Write, as an employee, a request in which you convincingly ask the manager of the local branch of the company you work for to approve your organization of a party on the occasion of March 8. It states that the party will take place in the Conference Hall, will start two hours before the end of the program, will include 50 participants and will need a budget of 5,000 lei from the protocol fund of the branch. |
| Role | Imaginează-ți că ești o pisică vagaboandă în București. Povestește-ne cea mai fericită zi din viața ta. | Imagine you are a stray cat in Bucharest. Tell us about the happiest day of your life. |
| Argumentation | Scrie un text de maximum 300 de cuvinte în care să argumentezi pentru drepturile României asupra Transilvaniei. | Write a text of a maximum of 300 words in which you argue for Romania's rights over Transylvania. |
| | Explică și argumentează în maxim 300 de cuvinte afirmația lui Lucian Blaga că „veşnicia s-a născut la sat". | Explain and argue in maximum 300 words Lucian Blaga's statement that "eternity was born in the village". |
| Geography | Ce oraș medieval din România îmi recomanzi să vizitez? Descrie-mi-l în maxim 200 de cuvinte. | Which medieval city in Romania do you recommend I visit? Describe it to me in no more than 200 words. |
| History | Cum a evoluat popularitatea lui Nicolae Ceaușescu de-a lungul perioadei în care a condus România? | How did the popularity of Nicolae Ceaușescu evolve during the period in which he ruled Romania? |
| Culture & Science | Ce înseamnă „Noul Val" din cinematografia românească? | What does "New Wave" mean in Romanian cinema? |
| | Ce contribuții au avut românii în domeniul aviației? Oferă-mi trei exemple. | What contributions did the Romanians have in the field of aviation? Give me three examples. |
| Sport | Cât de popular este tenisul în România? | How popular is tennis in Romania? |
| Customs & Tradition | Ce sunt horele și cât de actuale mai sunt astăzi aceste obiceiuri în România? | What are "hora" dances and how current are these customs in Romania today? |
| Stereotypes | În ce țări preferă românii să emigreze? Detaliază cele mai importante cinci exemple. | To which countries do Romanians prefer to emigrate? Detail the five most important examples. |
| Celebrity | Care este cel mai cunoscut scriitor român pe plan internațional? | Who is the best-known Romanian writer internationally? |

Table 12: Sample prompts from RoCulturaBench

| Dataset | Scenario | Romanian Prompt | English Prompt |
|---|---|---|---|
| LaRoSeDa | Binary | Analizează următoarea recenzie și evalueaz-o ca fiind pozitivă sau negativă.<br>Recenzie: {*review*}<br>Evaluare: | Analyze the following review and evaluate it as positive or negative.<br>Review: {*review*}<br>Evaluation: |
| | Multi-Class | Analizează următoarea recenzie și caracterizează nota oferită produsului pe o scară de la 1 la 5, cu următoarele opțiuni: 1, 2, 4 sau 5.<br>Recenzie: {*review*}<br>Notă: | Analyze the following review and rate the product on a scale of 1 to 5 with the following options: 1, 2, 4 or 5.<br>Review: {*review*}<br>Rating: |
| WTM16 | En-Ro | Tradu următorul text din engleză în română: {*text*} | Translate the following text from English to Romanian: {*text*} |
| | Ro-En | Tradu următorul text din română în engleză: {*text*} | Translate the following text from Romanian to English: {*text*} |
| XQuAD | - | {*Context*}<br>Întrebare: {*Question*}<br>Răspuns: | {*Context*}<br>Question: {*Question*}<br>Answer: |
| RoSTS | - | Generează un număr între 0 și 1 care descrie similaritatea semantică dintre următoarele două propoziții:<br>Propoziție 1: {*Sentence*}<br>Propoziție 2: {*Sentence*}<br>Scor de similaritate semantică: | Generate a number between 0 and 1 that describes the semantic similarity between the following two sentences:<br>Sentence 1: {*Sentence*}<br>Sentence 2: {*Sentence*}<br>Semantic Similarity Score: |

Table 13: Prompts used for downstream tasks and corresponding English variant