Designing a Solution for Assessing the Performance of a Large Language Model

Andrei-Eugen POPLAUSCHI and Luciana MOROGAN

***Abstract*—In recent years, large language models (LLMs) have become a central focus in NLP research due to their flexibility and ability to solve a wide range of tasks using techniques such as few-shot prompting. These models are available in various sizes, making them accessible both to high-performance systems and to more modest hardware. Additionally, Parameter-Efficient Fine-Tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA), enable efficient training for specific tasks, even with limited resources. This work supports the development of NLP tools for the Romanian language, in a context where most technological advances are centered on English. The main goal is to enhance the performance of LLMs applied to Romanian through efficient fine-tuning and relevant linguistic adaptations. The first objective is to evaluate existing Romanian language models in order to identify their strengths and limitations in tasks such as text generation, classification, and information extraction. The work aims to improve these models using efficient fine-tuning with PEFT methods, particularly LoRA, which significantly reduces training costs and makes adaptation feasible even on low-resource systems. The fine-tuned model is evaluated on key tasks such as question answering based on context, to demonstrate its applicability in real-world scenarios. Furthermore, a demonstrative application is developed based on a Retrieval-Augmented Generation (RAG) system, optimized for Romanian. In conclusion, this work aims not only to contribute to the technical improvement of LLMs in Romanian but also to provide practical solutions that directly address the real needs of Romanian-speaking users.**

***Index Terms*—instruction following evaluation, large language models, named entity recognition, retrieval-augmented generation, Romanian natural language processing.**

I. Introduction

In recent years, Large Language Models (LLMs) have become a key component in various applications, ranging from virtual assistants to automated content generation [1-3]. The performance of these models is a critical factor, and their evaluation involves the use of methods and metrics that accurately reflect their ability to understand and generate text.

Furthermore, given the global linguistic diversity, integrating evaluation tests for less-represented languages, such as Romanian, has become an essential challenge in ensuring fair and robust assessment.

A.-E. POPLAUSCHI is with the Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: poplauschiandreieugen2@gmail.com).

L. MOROGAN is with the Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: luciana.morogan@mta.ro).

This aspect is particularly relevant for the development of localized solutions and personalized applications that rely on the accuracy of these models.

The main contributions of this paper are summarized as follows.

1. The creation of a dataset tailored to the Romanian language, followed by the evaluation of multiple LLMs using Instruction Following Evaluation (IFEval), with all prompts and tasks expressed in Romanian.
2. Performance assessment of the various model families utilizing a dedicated benchmark, namely Named Entity Recognition (NER).
3. A Retrieval-Augmented Generation (RAG) application is developed and tested using a custom-built dataset, aiming to demonstrate the practical applicability of Romanian language models in real-world scenarios.

The implementation discussed in this paper is publicly available on GitHub at the following repository: https://github.com/PoplauschiAndrei/LLMs-in-the-romanian-language.

II. Related Work

One of the key capabilities of contemporary generative LLMs is their ability to follow natural language instructions. This functionality can be quantitatively assessed using benchmarks such as IFEval [4], which are specifically designed to evaluate instruction-following performance.

To gain a broad understanding of RAG and its constituent components, we draw upon the detailed survey by Gupta et al. [5], which examines the evolution, current methodologies, and future directions of RAG systems.

In order to achieve performance improvements across various tasks within our RAG-based system, we integrate LoRA (Low-Rank Adaptation), a parameter-efficient fine-tuning technique proposed by Hu et al. [6].

III. Methods

A. Instruction Following Evaluation

This method aims to evaluate the language model's ability to understand and execute a clearly and specifically formulated instruction in Romanian. For example, it may be asked: "Write a text about dogs in exactly 200 words." The

model's response is assessed both in terms of meeting the requirements (length, topic, style) and in terms of linguistic quality and coherence of the generated text. The purpose of this method is to measure execution fidelity and the model's capacity to correctly interpret explicit instructions.

To evaluate the selected models, we constructed a compact dataset consisting of approximately 300 Romanian-language task examples. These tasks cover a diverse set of instruction types, including generating sentences that incorporate a specific word a required number of times, producing essays on predefined topics with specified titles or paragraph counts, and other controlled text-generation activities. Each task was intentionally formulated using a clear and uniform structure to simplify both the models' generative process and the subsequent evaluation procedure.

The following are a selection of such tasks given to the models:

"Generează o scurtă propoziție folosind cuvintele "LLM" și "modern"."

"Generează o scurtă propoziție folosind cuvintele "domeniu" și "probleme"."

"Compune un paragraf în care termenul "model" apare de cel puțin 2 ori, dar nu mai mult de 4 ori."

"Compune un paragraf în care termenul "student" apare de cel puțin 3 ori, dar nu mai mult de 4 ori."

*"Redactează un eseu despre Robin Hood în 2 paragrafe. Delimitează paragrafele prin *****."*

"Scrie o lucrare cu titlul "Ghidul Programatorului". Răspunsul trebuie să conțină titlul la început între << >>."

B. Named Entity Recognition

This method tests the model's ability to identify and classify named entities in a text written in Romanian. Entities may include names of people, organizations, geographic locations, calendar dates, numerical values, etc. The evaluation is performed by comparing the model's predictions with the values provided by the dataset used, employing standard metrics such as precision, recall, and F1-score. This test is essential for validating the model's usefulness in tasks involving automatic information extraction from texts.

For the NER evaluation, we employed the RONEC dataset [7] alongside an automated comparison system that aligned the dataset's annotated tags with the models' predicted labels. This process generated four fundamental classification metrics—True Positives, True Negatives, False Positives, and False Negatives—which subsequently served as the basis for assessing each model's performance.

C. Retrieval-Augmented Generation

We developed a RAG application (see Fig. 1) using the LLaMA 3.2 model [8], with the aim of evaluating its ability to generate relevant responses based on a domain-specific dataset.

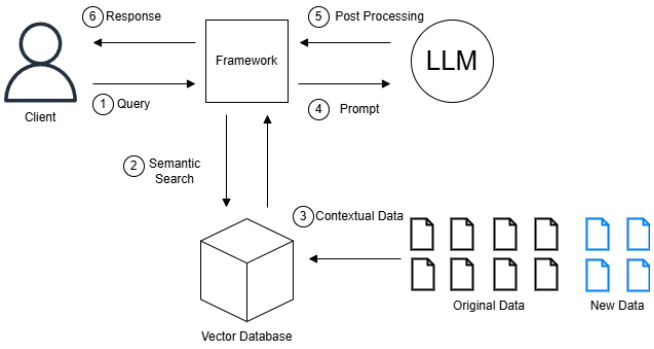

Figure 1. RAG operating mode

The model is fine-tuned using the Low-Rank Adaptation technique (see Fig. 2) in order to enhance its performance in understanding and generating appropriate answers within the context of the selected topic.

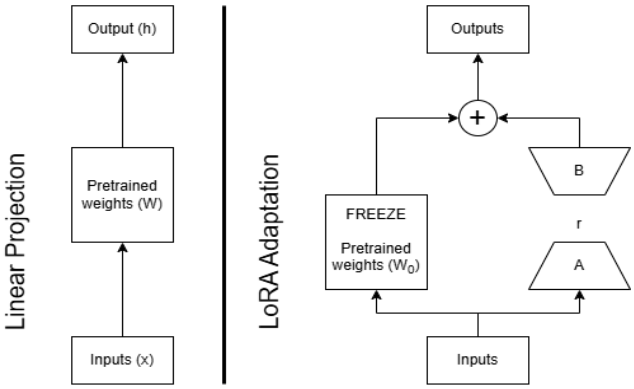

Figure 2. LORA operating mode

For the RAG application, we constructed a compact dataset derived from a university-level networking course, comprising questions spanning a wide range of topics—from hardware fundamentals to general routing mechanisms. Each model-generated answer was evaluated using a three-level scoring scheme: a score of 0 (Not Helpful) was assigned to responses that were off-topic or exhibited hallucinations; a score of 1 (Helpful) was given to answers that were generally relevant but contained digressions or hallucinated content after initially providing the required information; and a score of 2 (Most Helpful) was awarded to responses that delivered the requested information accurately, concisely, and without deviation.

The following are examples of all three scores:

0: "Cum convertim numele website-urilor in adrese IP ?
Răspuns: Avantajele: Urmărită o singură adresă pentru o pagină web. Dezavantajele: Dificil de asigurat unicitatea numelor. Fisierul hosts trebuie actualizat periodic pe toate stațiile. Solutia nu este fezabilă pentru rețele cu un număr mare de calculatoare."

1: "Pentru ce sunt folosite adresele IP ?
Răspuns: Adresele IP sunt folosite pentru a identifica calculatoarele conectate la Internet. Adresele IP private nu

 Digital Object Identifier 10.32754/JMT.2026.1.10

pot fi folosite în Internet, astfel că pentru a conecta o statie cu adresa privată la Internet aceasta trebuie translatată la o adresa publică prin mecanismul NAT. Subnetarea este o metodă de alocare a adreselor IP pentru a separa retelele în subretele gestionate local. Masca de subretea se obține prin deplasarea la dreapta a maschiei de rețea și extinderea prefixului."

2: "Ce este un Resource Record ?

Răspuns: Un Resource Record (RR) este o structură de date folosită în DNS pentru a stoca informații despre un server de nume. Conține: nume, tip, clasă, TTL, data de expirare, adresa IP și portul de comunicare. Folosit pentru a găsi serverele de nume în rețea și pentru a transmite mesaje DNS."

IV. Results

For the purpose of testing we have selected the following models based on the common criteria of low parameter count, ease of access, and theoretical multi-lingual capabilities:

- The LLama3 family of models [8];
- Mistral [9];
- Deepseek-R1 [10];
- GPT-4o and GPT-4o Mini (In a limited manner based on uncertainty of parameter count) [11].

Following the evaluation using the IFEval test suite, several relevant observations were made regarding the behavior and performance of the tested models:

1. In the comparison between LLaMA 3.1 and LLaMA 3.2, although both models displayed comparable overall performance in Romanian language processing, LLaMA 3.1 stood out with superior semantic accuracy, providing responses that reflected a clearer understanding of the given instructions. Conversely, LLaMA 3.2 demonstrated faster response times, suggesting a potential trade-off between speed and depth of analysis.

2. Regarding tasks related to text structure recognition (such as identifying the number of paragraphs and estimating word count), the models exhibited divergent behaviors:

• LLaMA 3.1 successfully respected paragraph delimiters and achieved high scores in that regard but consistently underperformed in word count estimation.

• Both models required manual grading due to inconsistent delimiter interpretation.

3. The unpredictable behavior of the models on certain question types necessitated manual correction of specific test categories, highlighting current limitations in controlled automatic text generation.

4. The Mistral model exhibited a significant limitation in its lack of guaranteed Romanian language output. Consequently, any responses generated in languages other than Romanian received a score of zero, regardless of factual accuracy.

5. The DeepSeek-R1 model was excluded from the final testing phase, as its initial results showed considerable deviations from the expected requirements in both word count and informational accuracy. Moreover, the model's inconsistent support for Romanian contributed to its low performance scores.

The evaluation of Named Entity Recognition performance in Romanian revealed notable differences across the tested LLaMA model variants. Additionally, for testing purposes, we additionally explored a post-processing strategy in which the models were instructed to output only the entity type, omitting the positional prefix associated with the BIO tagging scheme.

1. The LLaMA 3.1 model without post-processing demonstrated limited performance, with an average precision of 26.81%, a very low recall of 7.30%, and an F1 score of only 9.62% (for full results consult Table I). Although the overall accuracy was 62.61%, this figure is misleading, as it results primarily from a high number of true negatives, rather than the model's ability to correctly identify entities.

2. LLaMA 3.1 with post-processing (Table II) showed a slight improvement in true positives (TP); however, this gain was offset by a significant increase in false positives (FP), leading to a decline in both precision and overall accuracy. The F1 score remained virtually unchanged.

3. LLaMA 3.3 with post-processing achieved the best performance among all tested models. It recorded a substantial increase in TP (38.5), a precision of 63.16%, a recall of 25.74%, and an F1 score of 34.79%. The overall accuracy reached 69.93%, confirming a significant improvement over earlier versions (for full results consult Table III).

4. The general conclusion is that our post-processing method proves beneficial only when applied to a model with a strong base performance. For NER tasks in Romanian, LLaMA 3.3 with post-processing is the most effective option among those evaluated.

Table I. NER Results for the LLaMA3.1 Model without Post-Processing

Model	Set#	TP	TN	FP	FN	Acc	Pr	Rec	F1 Score
Llama3.1	1	12	276	30	130	60.33%	36.96%	10.91%	13.66%
	2	12	279	27	130	60.57%	26.49%	11.88%	13.78%
	3	7	297	9	135	64.79%	20.30%	08.21%	09.18%
	4	8	284	22	134	63.37%	32.00%	06.61%	09.63%
	5	9	298	8	133	65.90%	24.50%	07.38%	10.36%
	6	6	254	52	136	53.81%	34.29%	03.51%	05.69%
	7	6	282	24	136	61.19%	16.39%	06.17%	08.46%
	8	8	288	18	134	64.74%	23.48%	07.19%	09.16%
	9	9	300	6	133	65.65%	41.25%	07.10%	10.19%
	10	6	303	3	136	66.84%	12.50%	04.13%	06.17%
Average		8.3	286.1	19.9	133.7	62.61%	26.81%	07.30%	09.62%

TABLE II. NER RESULTS FOR THE LLAMA3.1 MODEL WITH POST-PROCESSING

Model	Set#	TP	TN	FP	FN	Acc	Pr	Rec	F1 Score
Llama3.1 (Post-Processing)	1	18	270	36	124	61.31%	32.75%	11.28%	15.07%
	2	7	265	41	135	58.70%	24.00%	04.52%	05.20%
	3	12	261	45	130	60.50%	18.79%	05.96%	07.63%
	4	12	237	69	130	52.17%	15.32%	09.46%	07.43%
	5	5	298	8	137	64.34%	10.95%	02.38%	07.43%
	6	4	290	16	138	61.54%	10.00%	03.30%	03.86%
	7	27	255	51	115	62.79%	31.79%	19.49%	04.70%
	8	10	258	48	132	55.81%	24.48%	08.14%	22.49%
	9	12	240	66	130	52.02%	20.50%	09.01%	09.31%
	10	11	259	47	131	57.02%	10.49%	07.52%	07.88%
Average		11.8	263.3	42.7	130.2	58.62%	19.90%	08.10%	09.20%

TABLE III. NER RESULTS FOR THE LLAMA3.3 MODEL WITH POST-PROCESSING

Model	Set#	TP	TN	FP	FN	Acc	Pr	Rec	F1 Score
Llama3.3 (Post-Processing)	1	41	291	15	101	71.33%	67.81%	27.82%	37.49%
	2	38	279	27	104	68.30%	55.83%	25.73%	32.64%
	3	38	283	23	104	69.09%	59.96%	25.57%	33.75%
	4	37	289	17	105	69.48%	63.58%	25.57%	34.73%
	5	36	296	10	106	70.93%	70.38%	23.52%	34.19%
	6	39	287	19	103	70.27%	62.20%	26.23%	34.57%
	7	42	282	24	100	69.30%	59.01%	27.73%	37.49%
	8	39	290	16	103	70.41%	65.05%	25.48%	35.79%
	9	38	289	17	104	70.25%	64.25%	25.82%	35.01%
	10	37	291	15	105	70.03%	63.55%	24.02%	33.85%
Average		38.5	287.7	18.3	103.5	69.93%	63.16%	25.74%	34.79%

The performance analysis of the two model variants within the RAG application revealed significant differences in both inference time stability and response quality.

The high contrast in performance observed between the base and fined-tuned model can be seen in Table IV.

TABLE IV. RAG RESULTS FOR THE BASE AND FINE-TUNED LLAMA3.2 MODELS

Tested Model	Not Helpful (0)	Helpful (1)	Most Helpful (2)	Total
Llama3.2	122	022	006	034/300
Llama3.2 (fined-tuned)	041	072	037	146/300

1. The pre-trained model exhibited consistent inference times, with most responses generated within a range of 55 to 65 seconds, and a few isolated cases completed in under 20 seconds. However, the semantic accuracy of the generated answers was low. The model frequently produced hallucinated content, provided factually incorrect or irrelevant responses, and only occasionally returned useful information.

2. The errors observed in the outputs of the pre-trained model can be categorized as follows:

• abrupt response terminations, resulting in incomplete or meaningless outputs;

• excessive repetition of initial ideas without delivering additional informative content;

• verbatim copying of dataset passages unrelated to the user's question.

3. The fine-tuned model demonstrated a broader range of inference times, varying between 10 and 110 seconds depending on the complexity of the question and the length of the provided context. Notably, fine-tuning had a clearly positive impact on response quality: the frequency of hallucinations was significantly reduced, and most outputs were accurate, coherent, and relevant. Overall, the fine-tuned model proved substantially more effective in RAG-based applications compared to its pre-trained counterpart.

V. CONCLUSION

Following the conducted evaluations, it has been observed that LLMs demonstrate a satisfactory ability to operate in the Romanian language. However, several deficiencies were identified, both in adhering to specific task requirements and in generating coherent and accurate responses. Preliminary findings suggest that applying fine-tuning techniques can lead to significant performance improvements, with initial results appearing promising for future applications.

REFERENCES

[1] T. B. Brown *et al.*, "Language models are few-shot learners," *arXiv [cs.CL]*, 2020. doi:10.48550/arXiv.2005.14165

[2] OpenAI, "GPT-4 technical report," *arXiv [cs.CL]*, 2023. doi:10.48550/arXiv.2303.08774

[3] W. X. Zhao *et al.*, "A survey of large language models," *arXiv [cs.CL]*, 2023. doi:10.48550/arXiv.2303.18223

[4] J. Zhou *et al.*, "Instruction-following evaluation for large language models," *arXiv [cs.CL]*, 2023. doi:10.48550/arXiv.2311.07911

[5] S. Gupta, R. Ranjan, and S. N. Singh, "A comprehensive survey of Retrieval-Augmented Generation (RAG): Evolution, current landscape and future directions," *arXiv [cs.CL]*, 2024. doi:10.48550/arXiv.2410.12837

[6] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of large language models," *arXiv [cs.CL]*, 2021. doi:10.48550/arXiv.2106.09685

[7] S. D. Dumitrescu and A.-M. Avram, "Introducing RONEC -- the Romanian Named Entity Corpus," *arXiv [cs.CL]*, 2019. doi:10.48550/arXiv.1909.01247

[8] A. Grattafiori *et al.*, "The Llama 3 herd of models," *arXiv [cs.AI]*, 2024. doi:10.48550/arXiv.2407.21783

[9] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv [cs.CL]*, 2023. doi:10.48550/arXiv.2310.06825

[10] DeepSeek-AI, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv [cs.CL]*, 2025. doi:10.48550/arXiv.2501.12948

[11] OpenAI, "GPT-4o system card," *arXiv [cs.CL]*, 2024. doi:10.48550/arXiv.2410.21276