Evaluation of Italian and English Small Language Models for Domain-based QA in Low-Resource Scenario

Irene Siragusa[1,*], Roberto Pirrone[1]

[1]Department of Engineering, University of Palermo, Palermo, 90128, Sicily, Italy

Abstract

Usage of open-source Large Language Models, which can be run locally, modified, fine-tuned, and queried without APIs that require data sharing, is required when dealing with sensitive or confidential information. In addition, suitable computational resources are needed to infer and fine-tune such models. The objective of this work is to assess the potentialities of Small Language Models in low-resource scenarios in which quantization may be required. In particular, the focus will be on the usage of these models in the context of the Italian and English languages from both a purely quantitative and resource-oriented evaluation, across two Question Answering data sets, a generic closed answer and a domain-based one with open answers.

Keywords

LLM, QA, Quantization, Fine-tuning,

1. Introduction

Generative Large Language Models (LLMs) are mainly oriented towards the paradigm "*the bigger the better*", involving both closed-source models such as GPT [1] Claude [2] and Gemini [3, 4], but also Llama (Llama 3.1 405B [5] or Llama 4 Maverick 400B [6]) and DeepSeek (DeepSeek R1 671B [7]) models. Despite the impressive capabilities of such models, in both textual and multimodal setup, significant issues arise when dealing with their size. In particular, higher computational resources are needed during the training phase, which is performed only once and asynchronously. The inference phase, on the other hand, despite requiring less computational resources, may result in being a bottleneck of the final distributed application, for which multi-currency and related GPU resources are needed. Pay-per-use APIs resolve all the computational aspects but lead to privacy-related issues. Applications that involve the use of Artificial Intelligence (AI) models as support systems in private companies or hospitals, where data is confidential or sensitive and any breaches must be avoided, should be compliant with those restrictive requirements and not allow sharing data with third parties.

To ensure these privacy-related issues, the focus of this work is on open-source models which can be trained locally and inferred in a low-resource scenario, both with full precision or in a quantization setup [8]. In doing this, the most recent Small Language Models (SLMs), released from late 2024 until April 2025, are considered, for which an instruction tuning phase was performed, involving both English and Italian as supported languages. This research led to the selection of models belonging to the following families, namely Qwen 3 [9], Gemma 3 [10], Phi 4 [11, 12] and Ministral [13], for which only the free models available below the 20B parameters are considered. Performances of these models were evaluated using both the full-precision and 8 /4 bit quantization scenarios. The evaluation was carried out with the generic benchmark MMLU [14, 15] and UniQA [16], a domain-specific Question Answer (QA) data set in the university domain. Both data sets cover English and Italian, and relative evaluations were performed in both languages. Statistics for the evaluation time and GPU used are also calculated. To further stress the potentialities of these models, the smallest ones were fine-tuned with two diverse strategies over the UniQA data set, and relative performances of both selected benchmarks have been analyzed.

Thus, the main contributions of this work can be summarized as follows.

1. Evaluation of open-source SLMs with MMLU benchmark and UniQA data set in different quantization scenarios, from both quantitative and computational perspective;
2. Fine-tuning with two proposed strategies over the UniQA data set;
3. Comprehensive evaluation of fine-tuned models over both MMLU and UniQA.

This work is organized as follows: an overview of fine-tuning strategies in the context of a low-resource scenario and of the selected open-source models and their principal characteristics are reported in Sections 2, and 3. The experimental setup along with the selected data sets and the proposed fine-tuning strategies are described in Sections 4 and 5. The results obtained are collected and

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.

irene.siragusa02@unipa.it (I. Siragusa);
roberto.pirrone@unipa.it (R. Pirrone)
0009-0005-8434-8729 (I. Siragusa); 0000-0001-9453-510X (R. Pirrone)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-11-30

discussed in Section 6, while the concluding remarks are drawn in Section 7.

2. Background

Fine-tuning pre-trained LLM in the context of domain and task adaptation involves strategies for both proper fine-tuning and the technique to reduce the overall fine-tuning computational cost, while keeping its effectiveness. Supervised Fine-Tuning (SFT) strategies are used for instruction tuning, domain, language, or task adaptation [17, 18, 19]. As a supervised method, both the input and the desired output are provided to the model, and, following a teacher forcing methodology, the model is forced to use the expected golden target token, even if the wrong one has been previously generated [20]. In the case of QA tasks, this consists of a question and associated answer and an optional context from which the answer should be derived.

Parameter-Efficient Fine-Tuning (PEFT) techniques are adopted in conjunction with SFT to speed up the fine-tuning phase and reduce computational resources required. In particular, those involve freezing, quantization, and Low-Rank Adaptation (LoRA) [21]. In freezing, only weights are actually trained in selected layers, while the rest are kept frozen. In the quantization strategy [8] the precision representation of the weights in the model is reduced from 32-bit to a 16-, 8-, or 4-bit representation. This technique can be used at both the training and inference time, thus decreasing the computational resources needed in terms of GPU. Lastly, LoRA [22] is one of the most used PEFT techniques where low-weight adapters, associated to selected layers, are actually trained, instead of the original ones. In doing this, the size of the trainable parameters is greatly decreased, and the computational resources needed for the fine-tuning process are reduced accordingly. In addition, those techniques can be combined to better fit computational constraints, such as in Quantized Low-Rank Adaptation (QLoRA) [23] in which quantization is applied along with LoRA during training.

For effective fine-tuning, models are trained, on average for a few training epochs, mainly ranging from 3 to 15, [24, 25, 26, 27], usually combining different PEFT strategies [28, 29].

3. Models

Capabilities around different tasks for closed-source and huge LLMs are well known, but in the context of real applications, usage of such models is impracticable. This can be mainly addressed to costly pay-per-use APIs, and to the sharing of private data to third parties that may lead to data breaches. Natural Language Processing (NLP) community is exploring not only the capabilities of larger [1, 2] and expert-based LLMs [6, 7, 4, 30], but also smaller models obtained through a distillation procedure from larger models [10], thus providing the general public with a valuable alternative.

Small Language Models are the focus of this research, which is limited to multilingual generative models with an explicit reference for supporting the Italian language, in addition to English. In particular, only models based on a transformer decoder-only architecture [31] and instruct fine-tuning are considered. Instruct models are capable of generating text given an instruction, thus making them suitable for the proposed evaluation scenario which includes closed and open QA tasks. In addition, as the increasing and faster development of newer models, only models which have been released from the last months of 2024 to April 2025 are examined. More in detail, we considered only models with less than 20B parameters, which have been sub-grouped as 4B, 8B, and 12B-14B models, to better evaluate their performance. The selected models are listed below along with their principal characteristics.

Gemma 3 [10] is a family of multimodal and multilingual models developed by Google DeepMind, co-designed with Gemini models [3, 4], with which they share the same tokenizer. A Grouped-Query Attention (GQA) mechanism [32] was used with post-norm and pre-norm with RMSNorm [33] and support for longer contexts. Gemma 3 models range from 1 to 27B parameters and were trained with a knowledge distillation strategy. In the context of this research, only the 4B and 12B versions are considered.

Ministral [13] is a model from the French company Mistral AI, it was released in the 3B and 8B parameter version. Ministral models are the newer version of Mistral 7B [34], which uses an interleaved sliding-window attention pattern to provide a faster, more computationally efficient, and low-latency solution at inference time. As the 3B version is not open-source, only the 8B version was considered in this work.

Phi 4 [11, 12] is a family of Microsoft models that showed impressive capabilities despite the reduced number of parameters, compared to other models. Higher performance of these models can be addressed to the three-stage training procedure and the data curation process, which involves a data decontamination process to the most used benchmarks. In addition, more variety in data, attention towards synthetic data for Chain of Thoughts (CoT) and reasoning capabilities, contributed in enhancing the overall behavior of these models. Phi 4 was released in its full version, which consists of 14B parameters and in its mini version with 3.8B parameters, which will be considered as a 4B model in the subsequent

analysis.

Qwen 3 [9] is a family of multilingual models released by the Chinese company Alibaba Cloud. Along with the large models of 30B and 235B parameters Mixture of Expert Models, smaller models have been released ranging from 1B to 32B parameters. Only models of 4B, 8B and 14B parameters are considered in this analysis. Great attention in Qwen 3 models was towards reasoning and CoT, both in training data selection and at inference time, in which the explicit thinking mode can be enabled or not.

4. Data sets

Three English and Italian data sets have been considered for evaluation purposes, two are closed QA data sets, and the other an open QA dataset. In the first case, the model is asked to answer with one of the provided answers, while in the second case a free text answer is expected. As closed QA, the general Massive Multitask Language Understanding (MMLU) task was selected in its English and Italian versions. From here on, the English version of MMLU will be referred to as MMLU-EN, and the Italian version as MMLU-IT, while MMLU will be used to refer to both splits. On the other hand, UniQA was selected as a domain-specific open answer QA data set in the university domain, available both in English and Italian.

MMLU-EN [14] is a generic benchmark task to evaluate the capabilities of LLMs after their training phase. It is a closed QA task involving 57 different subjects in STEM, humanities, and social science with diverse complexity ranging from elementary level to advanced and professional level. It consists of 14079 questions, and the models are queried with a 5-shot strategy in which 5 sample questions are provided for each subject. Accuracy is the proposed metric for performance evaluation.

MMLU-IT [15] is the translated version of the MMLU data set, which is also referenced in the Language Model Evaluation Harness framework [35]. Translation was obtained automatically using an *ad hoc* developed prompt for ChatGPT. No further checks have been conducted on the data set to evaluate its correctness in terms of translation.

UniQA [16] is a QA data set for the University domain that comprehends nearly 14k QA pairs and more than 1k documents, which serve as a context for the question. The data set has been generated in a semi-automated manner using the data retrieved from the website of the University of Palermo, covering information about the bachelor and master degree courses for the academic year 2024/2025. Data are natively both in Italian and English, i.e. no translation procedure was involved for developing the model. From here on, UniQA-EN will be used for the English split, UniQA-IT for the Italian one, and the general form UniQA will be used for both splits.

5. Experimental setup

Models in an out-of-the-box setup were tested with MMLU and UniQA data sets at different levels of quantization, namely in their base, 8-bit (Q8) and 4-bit (Q4) quantization [8]. These evaluations were performed to assess the different performances of quantized models versus their base version along with the effective computational resources involved, such as GPU memory and inference time. Quantization was performed with the usage of the bitsandbytes library[1] in combination with the transformers library [36] in both 8-bit and 4-bit quantization [37].

Regarding the MMLU-EN and MMLU-IT evaluation, we used the Language Model Evaluation Harness framework [35], in 5-shot setup, and considering the accuracy as the evaluation metric. Performance for the UniQA data set was obtained providing the following prompt, enriched with the target question and associated documents, following an in-context learning strategy [38].

> You are Unipa-GPT, the chatbot and virtual assistant of the University of Palermo.
>
> Provide an answer to the provided QUESTION concerning the University of Palermo, relying on the given DOCUMENTS
>
> If the question is in English, answer in English.
>
> If the question is in Italian, answer in Italian.
>
> QUESTION:
>
> `question`
>
> DOCUMENTS:
>
> `documents`

For UniQA, we used the default generation configurations suggested by the developers of the selected models. In particular, the thinking mode was disabled for Qwen 3, while the sampling strategy in the generation phase was disabled in the context of Gemma 3 models. Whenever the model was not able to generate and answer, the

[1] https://github.com/bitsandbytes-foundation/bitsandbytes

default empty answer has been considered as the generated one. As evaluation metric, BLEU [39], ROUGE [40], METEOR [41] and BERTScore [42], with the multilingual model XLM-RoBERTa Large [43] were calculated. Since the F1 BERTScore provides a more comprehensive evaluation of the meaning and significance of the generated answer, it was the only metric considered for evaluation purposes in the context of this work. In the Appendix, all the calculated metrics for the UniQA data set are reported for each inference configuration tested (Table 6).

5.1. Fine-tuning strategies

Only the smallest models, namely Gemma 3 4B, Phi 4 mini, and Qwen 3 4B, have been fine-tuned over the English and Italian training split of the UniQA data set. In particular, two different fine-tuning strategies have been proposed and used in this phase, namely *w/ docs* (with documents) and *w/o docs* (without documents). They differ in the arrangement of the training samples and the associated instruction prompt as reported in Table 1.

Table 1
Instruction prompts designed for fine-tuning w/ and w/o documents.

***w/ docs* prompt text**
You are Unipa-GPT, the chatbot and virtual assistant of the University of Palermo. Provide an answer to the provided QUESTION concerning the University of Palermo, relying on the given DOCUMENTS If the question is in English, answer in English. If the question is in Italian, answer in Italian. QUESTION: <QUESTION> DOCUMENTS: <DOCUMENTS>
***w/o docs* prompt text**
You are Unipa-GPT, the chatbot and virtual assistant of the University of Palermo. Provide an answer to the provided QUESTION concerning the University of Palermo. If the question is in English, answer in English. If the question is in Italian, answer in Italian. QUESTION: <QUESTION>

In w/ docs strategy, annotated documents were fed as input in the training sample, thus allowing the model to read the documents and force it to extract and re-paraphase the desired snippet in the document, containing the answer. On the other hand, in the w/o document strategy, no additional context was provided in the prompt, allowing the model to learn the QA pairs directly and, at inference time, to integrate the knowledge provided by the documents in a context-learning set-up [38].

Following the approaches described in Section 2, our choice was to perform a full fine-tuning limited to selected layers. A unique strategy was designed that is suitable for heterogeneous models with a different number of decoder layers. We fully fine-tuned only the last 25% of the decoder layers and the classification head, while freezing the remaining layers. The proposed strategy resulted in a valuable trade-off with PEFT techniques and full fine-tuning. In addition, this strategy meets the proposed research question in analyzing the impact of quantization at the inference phase and not during training.

The models have been trained for five epochs: a larger number of training epochs do not lead to significant improvement compared to the considered training data. A validation set was expunged from the training set with a 90:10 ratio, and it was used as a criterion to select the best model over the validation loss.

Inferences were run on a local machine on a single 48 GB NVIDIA RTX 6000 Ada Generation GPU (machine 1) and on a cluster with 1 NVIDIA A100 64 GB GPUs from the Leonardo supercomputer[2] via an ISCRA-C application (machine 2), while fine-tuning was executed on machine 2. Over the same machines, the occupied GPUs and inference time were monitored to simulate and provide an estimation of the required computational resources in the low-resource scenario.

6. Results

In Table 2 a comprehensive evaluation of the selected models is reported. Evaluations also include performances over bigger models such as Mistral Small [44], Llama 4 Scout Instruct [6], Claude 3.5 Sonnet [2] and GPT 4o Mini [38]. These models have been considered since their performance for both tasks was available from public leaderboards [45, 46, 47]. In this phase, no spot checks or roundtrip translations have been conducted to further investigate errors in the automatically translated MMLU-IT split, to assess whether some inference errors derive from actual model limitations or from translation artifacts.

The overall best results are achieved by Claude 3.5 Sonnet, followed by GPT 4o mini. Nevertheless, Phi 4 results in a valuable alternative since it reaches performances comparable to Mistral Small and is only 0.2

[2] https://leonardo-supercomputer.cineca.eu/it/home-it/

Table 2
Performance over MMLU-EN and MMLU-IT. The average performance over MMLU and the execution time in seconds for each sample are also reported. The bold values refer to the highest for each block, while the starred ones are the overall best. Runs have been performed on machine 2.

	Base inference				Q8 inference				Q4 inference			
Model	**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**	**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**	**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**
Gemma 3 4B	0.584	0.532	0.558	**0.22**	0.580	0.532	0.556	**0.14**	0.562	0.501	0.532	0.15
Phi 4 mini 4B	0.686	0.537	0.612	0.25	0.681	0.530	0.605	0.16	0.637	0.499	0.568	**0.07***
Qwen 3 4B	**0.701**	**0.651**	**0.676**	0.38	**0.702**	**0.651**	**0.676**	0.28	**0.665**	**0.604**	**0.635**	0.29
Ministral 8B	0.649	0.585	0.617	**0.10***	0.647	0.582	0.615	**0.17**	0.627	0.553	0.590	**0.13**
Qwen 3 8B	**0.749**	**0.708**	**0.729**	0.32	**0.747**	**0.705**	**0.726**	0.31	**0.728**	**0.669**	**0.698**	0.43
Gemma 3 12B	0.721	0.672	0.697	**0.20**	0.718	0.670	0.694	**0.10***	0.696	0.638	0.667	0.16
Phi 4 14B	**0.803**	**0.747**	**0.775**	0.24	**0.803***	**0.748***	**0.775***	0.24	**0.794***	**0.729***	**0.762***	**0.14**
Qwen 3 14B	0.788	0.738	0.763	0.28	0.784	0.729	0.756	0.27	0.776	0.722	0.749	0.39
Mistral Small 24B	0.806	0.758	0.782									
Llama 4 Scout	0.743	0.748	0.746									
Claude 3.5 Sonnet	0.790	**0.817***	**0.803***									
GPT 4o Mini	**0.820***	0.683	0.751									

Table 3
Performance over UniQA-EN and UniQA-IT as BERT-F1 score (B-F1). The average performance over UniQA, the execution time in seconds and GPU used in GB are also reported. The bold values refer to the highest for each block, while the starred ones are the overall best. Underlined results are the ones obtained over machine 2.

	Base inference					Q8 inference					Q4 inference				
	UniQA-EN	**UniQA-IT**	**UniQA**			**UniQA-EN**	**UniQA-IT**	**UniQA**			**UniQA-EN**	**UniQA-IT**	**UniQA**		
Model	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**
Gemma 3 4B	0.870	**0.896***	**0.883**	11.1 ± 5.0	17.8 ± 0.1	0.748	0.730	0.739	58.7 ± 26.7	5.2 ± 0.0	0.001	0.001	0.001	19.6 ± 1.1	3.8 ± 0.1
Phi 4 mini 4B	0.877	0.881	0.879	**8.4 ± 6.3***	**15.4 ± 0.0***	**0.877**	**0.881**	**0.879**	**15.6 ± 9.6***	4.5 ± 0.0	**0.874**	**0.876**	**0.875**	**12.5 ± 9.4**	3.1 ± 0.0
Qwen 3 4B	**0.879**	0.879	0.879	9.9 ± 7.6	16.2 ± 0.0	**0.877**	0.880	0.878	37.8 ± 50.0	**4.4 ± 0.0***	0.868	0.870	0.869	22.1 ± 10.6	**2.9 ± 0.0***
Ministral 8B	**0.878**	**0.887**	**0.882**	15.6 ± 10.9	**32.1 ± 0.0**	**0.877**	**0.887***	**0.882***	48.3 ± 36.0	**9.1 ± 0.0**	**0.882***	**0.892***	**0.887***	13.7 ± 9.4	**6.1 ± 0.2**
Qwen 3 8B	0.872	0.877	0.874	**14.0 ± 9.9**	32.8 ± 0.0	0.872	0.878	0.875	**30.6 ± 13.6**	9.5 ± 0.0	0.87	0.875	0.872	**9.7 ± 5.2**	6.4 ± 0.0
Gemma 3 12B	**0.883***	**0.888**	**0.886***	15.3 ± 5.6	**50.0 ± 0.1**	0.668	0.633	0.650	109.4 ± 22.4	**14.4 ± 0.7**	0.000	0.000	0.000	53.3 ± 2.0	**9.0 ± 0.4**
Phi 4 14B	0.842	0.743	0.793	**14.6 ± 6.1**	58.7 ± 0.0	**0.879***	**0.884**	**0.881**	**21.4 ± 6347.0**	15.7 ± 0.0	0.869	**0.882**	**0.876**	**8.1 ± 3448.5***	9.8 ± 0.0
Qwen 3 14B	0.824	0.841	0.833	16.0 ± 5009.1	59.1 ± 0.0	0.872	0.881	0.877	38.2 ± 16.6	16.4 ± 0.0	**0.870**	0.879	0.875	19.1 ± 8.0	10.6 ± 0.0

points below GPT in MMLU-EN. As for MMLU-IT, scores tend to be lower compared to the English split, and again Phi results the best. With reference to smaller models, Qwen 3 in its 4B and 8B versions outperforms other models in MMLU tasks, while showing a significant average inference time, compared with the competitors. Performance generally exhibits a decrease in quantized models. The decrease is significant in the case of Q4, while the average inference time for each question is decreasing for Q8 and tends to increase in the case of Q4, especially for Qwen.

Generally speaking, the quantization procedure at inference time can increase the answer time due to the additional computation required for quantization [37]. This behavior is highly emphasized with the UniQA evaluation, where the input provided to the models can be significantly longer compared to MMLU samples. The results for UniQA are reported in Table 3, together with the average GPU occupied for each inference. To better compare obtained results, the standard deviation over the average inference time and GPU usage is also reported. Note that the average inference time is reported in seconds, while the GPU usage in GB: associated standard deviation follows the same scale, and, in the GPU case, the majority results 0.0 since the corresponding variation is lower than 0.1 GB.

In full precision inference, best results are assessed by Gemma 3 models reaching a BERT-F1 score of 0.88 on average in both 4B and 12B versions, also surpassing larger 14B models. In this context, the performances of Gemma 3 4B are much more interesting from a computational perspective, since it is 64% smaller than the 12B version, reaching comparable performances. The smallest model in this set-up is Phi 4B mini, which occupies less than 16GB and reaches the smallest inference time, which is desired in context of real-time applications. Regarding quantized inferences, GPU values decrease by 70% and 80% for Q8 and Q4, respectively, compared to models inferred with full precision. In terms of inference time, significant increases are found in Q8, while a reduction is found in Q4, which is mainly related to the quantization strategy adopted by bitsandbytes [37]. Overall, for both performance and computational resource usage, Phi and Ministral are the best models which benefit from quantization, and keep comparable performances over the selected benchmarks, despite a slight decrease. The worst performances are assessed by Gemma models which deeply suffer the quantization procedure that leads to output empty string (Q4) or meaningless output in not desired languages (Q8).

In contrast with the MMLU case, in which a slight discrepancy can be found between the English and

Table 4
Performance over MMLU and UniQA for fine-tuned models without docs strategy. The average performance and execution time in seconds and GPU used in GB are also reported. Bold values refer to the highest ones. Runs have been performed on machine 1.

	Base inference					Q8 inference					Q4 inference				
Model	**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**		**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**		**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**	
Gemma 3 4B	0.579	0.518	0.548	**0.22**		0.574	0.519	0.547	0.20		0.552	0.487	0.519	0.15	
Phi 4 mini 4B	0.672	0.519	0.596	0.24		0.666	0.508	0.587	**0.18**		0.631	0.480	0.555	**0.08**	
Qwen 3 4B	**0.678**	**0.620**	**0.649**	0.34		**0.674**	**0.614**	**0.644**	0.26		**0.648**	**0.577**	**0.612**	0.28	
	UniQA-EN	**UniQA-IT**	**UniQA**			**UniQA-EN**	**UniQA-IT**	**UniQA**			**UniQA-EN**	**UniQA-IT**	**UniQA**		
Model	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**
Gemma 3 4B	0.929	0.881	0.905	4.5 ± 4.4	17.7 ± 0.3	0.739	0.732	0.736	62.6 ± 50.9	5.3 ± 0.1	0.000	0.000	0.000	19.4 ± 1.6	3.7 ± 0.1
Phi 4 mini 4B	**0.959**	**0.926**	**0.942**	4.1 ± 3.9	**15.4 ± 0.0**	0.959	**0.944**	**0.952**	**4.7 ± 4.6**	4.5 ± 0.0	0.940	0.898	0.919	3.9 ± 3.8	3.4 ± 0.0
Qwen 3 4B	0.957	0.925	0.941	**4.1 ± 2.3**	16.2 ± 0.0	**0.960**	0.942	0.951	7.4 ± 4.6	**4.4 ± 0.0**	**0.963**	**0.940**	**0.950**	**3.0 ± 2.0**	**2.9 ± 0.0**

Table 5
Performance over MMLU and UniQA for fine-tuned models with docs strategy. The average performance and execution time in seconds and GPU used in GB are also reported. Bold values refer to the highest ones. Runs have been performed on machine 1.

	Base inference					Q8 inference					Q4 inference				
Model	**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**		**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**		**↑ MMLU-EN**	**↑ MMLU-IT**	**↑ MMLU**	**↓ TIME**	
Gemma 3 4B	0.581	0.522	0.552	**0.54**		0.578	0.521	0.550	0.25		0.553	0.490	0.522	0.18	
Phi 4 mini 4B	**0.680**	0.535	0.607	0.57		0.671	0.527	0.599	**0.16**		0.631	0.492	0.562	**0.10**	
Qwen 3 4B	0.675	**0.623**	**0.649**	0.64		**0.673**	**0.620**	**0.646**	0.41		**0.651**	**0.583**	**0.617**	0.37	
	UniQA-EN	**UniQA-IT**	**UniQA**			**UniQA-EN**	**UniQA-IT**	**UniQA**			**UniQA-EN**	**UniQA-IT**	**UniQA**		
Model	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**	**↑ B-F1**	**↑ B-F1**	**↑ B-F1**	**↓ TIME**	**↓ GPU**
Gemma 3 4B	0.933	0.917	0.925	5.3 ± 4.7	17.8 ± 0.3	0.752	0.665	0.708	57.5 ± 44.5	5.3 ± 0.2	0.000	0.000	0.000	19.3 ± 3.0	3.7 ± 0.1
Phi 4 mini 4B	0.953	**0.927**	0.940	**4.8 ± 4.8**	**15.4 ± 0.0**	0.951	**0.942**	0.946	**5.6 ± 5.7**	4.5 ± 0.0	0.930	0.911	0.920	4.4 ± 4.3	3.1 ± 0.0
Qwen 3 4B	**0.965**	0.922	**0.944**	4.8 ± 4.4	16.2 ± 0.0	**0.964**	0.938	**0.951**	7.1 ± 4.5	**4.5 ± 0.0**	**0.957**	**0.925**	**0.941**	**3.7 ± 3.4**	**2.9 ± 0.0**

Italian split, in the UniQA case, all performance places on the same level, and, in some cases, slightly towards the Italian split. This performance can be explained through an analysis on the data set, in which the presence of the context can guide the model more effectively in generating the desired answer, and in the language-related characteristics and understandings.

In Tables 4 and 5 the results over both benchmarks are reported using the two proposed fine-tuning strategies, with and without documents.

No improvements are found after the fine-tuning phase with the two proposed strategies in terms of performance on the MMLU benchmarks. Models fine-tuned with the w/ docs strategy tend to better maintain the performance obtained by the base models. These results show that the fine-tuning on a specific task did not lead to a degradation in performance in a generic benchmark and that the generalization performance of the considered LLM is maintained. This is mainly due to the light fine-tuning strategy adopted, which does not cause the model to overfit.

Regarding UniQA performance, both strategies have been shown to be successful since overall performance for the BERT F1 score increased. More specifically, better results are obtained in the case of w/o docs strategy, both from evaluation metrics and for average inference time, which is reduced. Improvements are found in both the base and quantized inferences. As in the without fine-tuning inference, the average time for quantized models deeply penalized Gemma 3 4B, while Qwen 3 4B trained with a w/ docs strategy, resulted in being the overall best model both in MMLU and UniQA benchmarks. Qwen 3 4B, in fact, better maintained the same level of performance across the different quantization levels. In addition, the w/o docs fine-tuning strategy was crucial to improve capabilities for Phi 4 mini 4B, in particular in the base and Q8 quantized inference. A general speed-up in performances is found in fine-tuned models over UniQA benchmark, while no improvements are found in Gemma 3 4B in Q4 setup, where performances are kept low.

The results obtained show that recent progress in developing multilingual LLMs provides the opportunity to use a valuable out-of-the-box model, also for domain-specific tasks with appropriate prompt engineering. In addition, the two proposed fine-tuning strategies, coupled with an overall light training phase as for the number of epochs, trainable layers, and consequently the resources needed, results crucial to improve capabilities of the SLMs under consideration, as for Phi 4 mini 4B and Qwen 3 4B. Those models trained in a target domain for a desired QA task of interest were able to outperform models three times larger in size, requiring on-budget resources. In general, both models should be considered as a valuable alternative to develop a custom LLM in a low-resource scenario. Phi tend to outperform after a w/o docs fine-tuning in terms of BERT score. On the other hand, Qwen presents strong performance in both traditional metrics such as the BLEU, ROUGE, and Meteor scores (Table 6) with both a fine-tuning strategy and

different quantization. Depending on the actual computational resources available, Phi is preferred, since it is smaller compared to Qwen. Despite metrics being really close to each other, the w/o docs training strategy is the best and the fastest one in the training phase.

7. Conclusions

In this work, we evaluated the recent open-source instruction-tuned multilingual Small Language Models belonging to different families with a focus on their performances upon a base inference and after a Q8 and Q4 quantization. In particular, both closed and open answer QA tasks were analyzed in Italian and English. Performances were evaluated from a quantitative perspective with the general MMLU benchmark and UniQA, a QA data set based on a specific domain for which relevant documents are associated to each question.

The results show that among the largest models under evaluation, Phi 4 14B almost reached Claude 3.5 Sonnet and GPT 4o Mini in the MMLU benchmark, while Gemma 3 14B obtained interesting performances when inferred with full precision using UniQA. Among the smaller models, Qwen 3 4B and Phi 4 mini 4B were the most promising ones: both models better scale in terms of performance after Q8 and Q4 in selected benchmarks.

In addition, two fine-tuning strategies were proposed for the last 25% layers and the classification head of the smaller models using the training split of the UniQA data set. The results proved that Qwen 3 4B benefits the most of the training when evaluated over UniQA, while maintaining general good performance in the MMLU task. Such considerations together with the flexibility towards quantization and smaller inference time make Qwen 3 4B a valuable model to implement custom LLM-based applications in a low-context scenario after a suitable fine-tuning phase.

More tests are needed to evaluate the performance of the investigated models from a qualitative perspective. More in detail, additional tests will be conducted to simulate a real-case scenario, involving both human evaluation of the quality of the provided answers and truly open-ended QA in the domain of interest.

Acknowledgments

We thank Giampiero Barbaro, who contributed in developing the training strategy in the early stages of this work, as part of his master thesis. This work is supported by the cup project J73C24000070007, "CAESAR" (Cognitive evolution in Ai: Explainable and Self-Aware Robots through multimodal data processing). The works presented were partially developed on the Leonardo supercomputer with the support of the CINECA-Italian Super Computing Resource Allocation class C project IscrC_DOCVLM2 (HP10C97VNN).

Declaration on Generative AI

During the preparation of this work, the authors used Writefull for grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] OpenAI, GPT-4o System Card, arXiv preprint arXiv:2410.21276 (2024).

[2] Antropic, The Claude 3 Model Family: Opus, Sonnet, Haiku, 2024.

[3] GeminiTeam, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., Gemini: A Family of Highly Capable Multimodal Models, 2024.

[4] Google DeepMind, Gemini 2.5: Our most intelligent AI model, 2025. blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

[5] LlamaTeam, The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783 (2024).

[6] LlamaTeam, The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

[7] DeepSeek-AI, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv preprint arXiv:2501.12948 (2025).

[8] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko, Quantization and training of neural networks for efficient integer-arithmetic-only inference, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[9] QwenTeam, Qwen3 Technical Report, arXiv preprint arXiv:2505.09388 (2025).

[10] GemmaTeam, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, et al., Gemma 3 Technical Report, arXiv preprint arXiv:2503.19786 (2025).

[11] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, et al., Phi-4 Technical Report, arXiv preprint arXiv:2412.08905 (2024).

[12] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, et al., Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs, arXiv preprint arXiv:2503.01743 (2025).

[13] MistralAITeam, Un Ministral, des Ministraux, 2024. https://mistral.ai/news/ministraux.

[14] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, arXiv preprint arXiv:2009.03300 (2021).

[15] V. D. Lai, C. V. Nguyen, N. T. Ngo, T. Nguyen, F. Dernoncourt, R. A. Rossi, T. H. Nguyen, Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback, arXiv preprint arXiv:2307.16039 (2023).

[16] I. Siragusa, R. Pirrone, UniQA: an italian and english question-answering data set based on educational documents, Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024) (2024).

[17] Alpaca-LoRA, https://github.com/tloen/alpaca-lora, 2023.

[18] C. Xu, D. Guo, N. Duan, J. McAuley, Baize: An open-source chat model with parameter-efficient tuning on self-chat data, arXiv preprint arXiv:2304.01196 (2023).

[19] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, 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), ELRA and ICCL, Torino, Italia, 2024, pp. 4343–4355. URL: https://aclanthology.org/2024.lrec-main.388/.

[20] D. Jurafsky, J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 2025.

[21] V. Lialin, V. Deshpande, X. Yao, A. Rumshisky, Scaling down to scale up: A guide to parameter-efficient fine-tuning, arXiv preprint 2303.15647 (2024).

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, in: International Conference on Learning Representations, 2022.

[23] Dettmers and Tim and Artidoro Pagnoni and Ari Holtzman and Luke Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, arXiv preprint arXiv:2305.14314 (2023).

[24] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[25] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024.

[26] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, O. Levy, LIMA: Less Is More for Alignment, arXiv preprint arXiv:2305.11206 (2023).

[27] W. Lu, R. K. Luu, M. J. Buehler, Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities, 2024.

[28] A. Afzal, R. Chalumattu, F. Matthes, L. Mascarell, AdaptEval: Evaluating Large Language Models on Domain Adaptation for Text Summarization, in: Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), 2024, pp. 76–85.

[29] J. Zheng, H. Hong, F. Liu, X. Wang, J. Su, Y. Liang, S. Wu, Dragft: Adapting large language models with dictionary and retrieval augmented fine-tuning for domain-specific machine translation, arXiv preprint arXiv:2402.15061 (2024).

[30] MistralAITeam, Large Enough, 2024. https://mistral.ai/news/mistral-large-2407.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, 2017.

[32] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, S. Sanghai, GQA: Training generalized multi-query transformer models from multi-head checkpoints, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 4895–4901. URL: https://aclanthology.org/2023.emnlp-main.298/. doi:10.18653/v1/2023.emnlp-main.298.

[33] B. Zhang, R. Sennrich, Root mean square layer normalization, Curran Associates Inc., Red Hook, NY, USA, 2019.

[34] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, arXiv preprint arXiv:2310.06825 (2023).

[35] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2024. URL: https://zenodo.org/records/12608602. doi:10.5281/zenodo.12608602.

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, et al., Transformers: State-of-the-Art Natural Language Processing, in: Proceedings

of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[37] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022. URL: https://arxiv.org/abs/2208.07339. arXiv:2208.07339.
[38] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-Shot Learners, Advances in neural information processing systems (2020).
[39] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002.
[40] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, 2004.
[41] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005.
[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, arXiv preprint arXiv:1904.09675 (2020).
[43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, 2019.
[44] MistralAITeam, Mistral Small 3.1, 2025. https://mistral.ai/news/mistral-small-3-1.
[45] Multi-task Language Understanding on MML, https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu, 2021.
[46] Y. Zhou, Y. Sakai, Y. Zhou, H. Li, J. Geng, Q. Li, W. Li, Y. Lin, A. Way, Z. Li, Z. Wan, D. Wu, W. Lai, B. Zeng, Multilingual MMLU Benchmark Leaderboard, 2024. https://huggingface.co/spaces/StarscreamDeceptions/Multilingual-MMLU-Benchmark-Leaderboard.
[47] Classifica generale degli LLM italiani, https://huggingface.co/spaces/mii-llm/open_ita_llm_leaderboard, 2024.

A. Evaluation metrics

In Table 6, are reported the full calculated metrics over the UniQA data set in different quantization and fine-tuning strategies.

Table 6
Overview of the calculated metrics in the UniQA-EN and UniQA-IT split. BERT-prec and BERT-rec stands for BERT precision and recall scores, respectively, while FT and QTN refers to the fine-tuning and quantization strategy adopted. Average performance and execution time is reported in seconds, while the GPU used in GB. Bold values are the higher ones for each block, while starred ones the overall best.

UNIQA											
Model	**FT**	**QTN**	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-Lsum**	**Meteor**	**BERT-prec**	**BERT-rec**	**BERT-F1**
Gemma 3 4B			0.124	0.393	0.240	0.304	0.317	0.344	**0.882**	0.885	**0.883**
Phi 4 4B			**0.147**	0.399	0.238	0.298	0.307	0.408	0.870	0.889	0.879
Qwen 3 4B			0.134	**0.419**	**0.261**	**0.331**	**0.354**	**0.411**	0.866	**0.894**	0.879
Gemma 3 4B		Q8	0.000	0.017	0.000	0.014	0.014	0.016	0.728	0.753	0.739
Phi 4 4B		Q8	**0.142**	0.393	0.235	0.292	0.301	0.408	**0.869**	0.890	**0.879**
Qwen 3 4B		Q8	0.129	**0.415**	**0.258**	**0.328**	**0.355**	**0.409**	0.863	**0.895**	0.878
Gemma 3 4B		Q4	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001
Phi 4 4B		Q4	**0.148**	**0.404**	**0.240**	**0.297**	0.313	**0.396**	**0.868**	0.884	**0.875**
Qwen 3 4B		Q4	0.094	0.363	0.222	0.286	**0.319**	0.370	0.850	**0.890**	0.869
Gemma 3 4B	w/o docs		0.404	0.657	0.580	0.619	0.621	0.628	0.910	0.901	0.905
Phi 4 4B	w/o docs		0.524	0.793	0.735	0.771	0.778	0.789	**0.941**	**0.944**	**0.942**
Qwen 3 4B	w/o docs		**0.586**	**0.821**	**0.770**	**0.802**	**0.805**	**0.819**	0.939	0.943	0.941
Gemma 3 4B	w/o docs	Q8	0.001	0.034	0.000	0.025	0.027	0.032	0.722	0.752	0.736
Phi 4 4B	w/o docs	Q8	0.536	0.810	0.755	0.791	0.796	0.807	**0.950**	**0.954***	**0.952***
Qwen 3 4B	w/o docs	Q8	**0.598**	**0.829**	**0.778**	**0.809**	**0.813**	**0.828***	0.949	0.954	0.951
Gemma 3 4B	w/o docs	Q4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Phi 4 4B	w/o docs	Q4	0.396	0.649	0.576	0.614	0.619	0.655	0.916	0.922	0.919
Qwen 3 4B	w/o docs	Q4	**0.608**	**0.835***	**0.794**	**0.821***	**0.825***	**0.826**	**0.951**	**0.953**	**0.952**
Gemma 3 4B	w/ docs		0.543	0.760	0.720	0.741	0.745	0.750	0.923	0.927	0.925
Phi 4 4B	w/ docs		0.566	0.778	0.743	0.764	0.768	0.771	0.938	0.941	0.940
Qwen 3 4B	w/ docs		**0.607**	**0.831**	**0.801***	**0.816**	**0.819**	**0.815**	**0.945**	**0.942**	**0.944**
Gemma 3 4B	w/ docs	Q8	0.001	0.050	0.002	0.035	0.041	0.046	0.692	0.727	0.708
Phi 4 4B	w/ docs	Q8	0.564	0.776	0.740	0.762	0.766	0.771	0.944	0.949	0.946
Qwen 3 4B	w/ docs	Q8	**0.611***	**0.832**	**0.800**	**0.816**	**0.819**	**0.814**	**0.953***	**0.950**	**0.951**
Gemma 3 4B	w/ docs	Q4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Phi 4 4B	w/ docs	Q4	0.431	0.662	0.608	0.641	0.647	0.652	0.918	0.923	0.920
Qwen 3 4B	w/ docs	Q4	**0.546**	**0.782**	**0.735**	**0.754**	**0.757**	**0.766**	**0.943**	**0.939**	**0.941**
Ministral 8B			**0.120**	**0.422**	**0.263**	**0.329**	0.356	**0.425**	**0.870**	**0.896**	**0.882**
Qwen 3 8B			0.116	0.416	0.252	0.328	**0.360**	0.405	0.856	0.894	0.874
Ministral 8B		Q8	**0.120**	**0.423**	**0.263**	**0.330**	0.358	**0.426**	**0.870**	**0.896**	**0.882**
Qwen 3 8B		Q8	**0.120**	0.418	0.253	0.329	**0.363**	0.406	0.857	0.894	0.875
Ministral 8B		Q4	**0.142**	**0.442**	**0.284**	**0.347**	**0.366**	**0.440**	**0.879**	**0.895**	**0.887**
Qwen 3 8B		Q4	0.107	0.415	0.246	0.322	0.351	0.383	0.857	0.888	0.872
Gemma 3 12B			**0.155**	**0.500**	**0.304**	**0.389**	**0.431**	**0.453**	**0.868**	**0.905**	**0.886**
Phi 4 14B			0.122	0.415	0.241	0.303	0.345	0.405	0.778	0.809	0.793
Qwen 3 14B			0.117	0.399	0.253	0.324	0.341	0.398	0.818	0.848	0.833
Gemma 3 12B		Q8	0.000	0.002	0.000	0.002	0.002	0.001	0.646	0.656	0.650
Phi 4 14B		Q8	**0.125**	**0.449**	0.262	0.332	**0.378**	**0.443**	**0.865**	**0.899**	**0.881**
Qwen 3 14B		Q8	0.123	0.421	**0.268**	**0.342**	0.365	0.421	0.860	0.895	0.877
Gemma 3 12B		Q4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Phi 4 14B		Q4	**0.118**	**0.434**	**0.252**	**0.320**	**0.364**	**0.432**	**0.860**	**0.893**	**0.876**
Qwen 3 14B		Q4	0.110	0.397	**0.252**	**0.320**	0.337	0.402	0.859	0.892	0.875

Declaration on Generative AI

During the preparation of this work, the author(s) used Other and Writefull in order to: Improve writing style and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.