A study on the soundness of closed-ended evaluation of Large Language Models adapted to the Italian language

Elio Musacchio[1,2,*], Lucia Siciliani[1,*], Pierpaolo Basile[1,*], Edoardo Michielon[3], Marco Pasqualini[3], Asia Beatrice Uboldi[3] and Giovanni Semeraro[1]

[1]Department of Computer Science, University of Bari Aldo Moro, Italy
[2]National PhD in Artificial Intelligence, University of Pisa, Italy
[3]Fastweb SpA, Milan, Italy

Abstract

With the rising interest in Large Language Models, deep architectures capable of solving a wide range of Natural Language Generation tasks, an increasing number of open weights architectures have been developed and released online. In contrast with older architectures, which were aimed at solving specific linguistic assignments, Large Language Models have shown outstanding capabilities in solving several tasks at once, raising the question of whether they can truly comprehend natural language. Nevertheless, evaluating this kind of capability is far from easy. One of the proposed solutions so far is using benchmarks that combine various types of tasks. This approach is based on the premise that achieving good performance in each of these individual tasks can imply having developed a model capable of understanding language. However, while this assumption is not incorrect, it is evident that it is not sufficient, and the evaluation of Large Language Models still remains an open challenge. In this paper, we conduct a study aimed at highlighting the potential and limitations of current datasets and how a new evaluation setting applied to language-adapted Large Language Models may provide more insight than traditional approaches.

Keywords

Large Language Models, Natural Language Processing, Evaluation, Benchmark

1. Introduction

Large Language Models (LLMs) are models based on the Transformer architecture capable of solving a wide variety of *Natural Language Generation* (NLG) tasks, even those not encountered during training, due to their extensive training and large number of parameters. Thanks to their remarkable skills, interest in LLMs is now at its climax, resulting in a proliferation of open-weight models (e.g. LLaMA, Mistral, and many others). Among the several challenges related to the development of LLMs, one of the most critical is their evaluation [1]. One approach to tackle this issue has been to build benchmarks that collect different datasets, with the aim of obtaining a more comprehensive evaluation of the model's overall capabilities. Currently, there is a leaderboard[1] [2] which keeps track of the capabilities of openly available LLMs. Specifically, the models are tested on six tasks that span different abilities a language model should have, e.g. reasoning or text completion. Regarding their reasoning abilities, the models are tested by solving *closed-ended* tasks. Specifically, multiple-choice question answering tasks are provided, where a question is given with a list of possible alternatives associated with an identifier (a letter, a number, and so on). Intuitively, since the model has also been pre-trained on *closed-ended* question-answering data, it should be able to generalize and understand the correct choice out of the available ones. Furthermore, rather than generating the output directly, the probabilities learned by the model are studied, using log-likelihood to assess which option is more likely to be correct. For the English language, this evaluation methodology has been a standard approach to assess the capabilities of LLMs. However, when adapting a model to a new language, due to the low amount of non-English data that has been used to pre-train such models, this methodology may not be as sound. The model only has to generate the correct option identifier, therefore this is not really testing the ability of the model of generating high-quality text in another language. The goal of this work is to understand whether a new evaluation setting applied to language-adapted LLMs may give more insight than the traditional approach. Therefore, our contributions are the following:

- We test two evaluation settings for language-adapted LLMs changing the structure of *closed-*

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
elio.musacchio@uniba.it (E. Musacchio); lucia.siciliani@uniba.it (L. Siciliani); pierpaolo.basile@uniba.it (P. Basile); edoardo.michielon@consulenti.fastweb.it (E. Michielon); marco.pasqualini@consulenti.fastweb.it (M. Pasqualini); asiabeatrice.uboldi@consulenti.fastweb.it (A. B. Uboldi); giovanni.semeraro@uniba.it (G. Semeraro)
0009-0006-9670-9998 (E. Musacchio); 0000-0002-1438-280X (L. Siciliani); 0000-0002-0545-1105 (P. Basile); 0000-0001-6883-1853 (G. Semeraro)

[1]https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

ended question answering tasks;
- We evaluate the performance of state-of-the-art models on these settings;
- We study the sensitivity that the models have for the input prompt.

2. Related Works

Language Model evaluation has been a research focus ever since the first Decoder-only models, which were designed for *natural language generation*.

One of the most remarkable skills regarding LLMs reasoning has been *in-context learning*. In particular, *few-shot learning* has been increasingly used. The idea is that providing examples of input-output in the model prompt should affect positively the generation process [3].

There are multiple leaderboards which evaluate open LLMs on non-English languages, e.g. *Open PL LLM Leaderboard* [4] for Polish or *Open KO LLM Leaderboard* [5] for Korean. These leaderboards are often based on the *lm evaluation harness* framework [6], which has been a milestone in the evaluation of LLMs. LLM evaluation can also depend on the topic at hand. There are some works which focus on mathematical reasoning [7] as well as factuality [8].

These evaluation settings often rely on *closed-ended* tasks, specifically multiple-choice question answering. The idea is to calculate the log-likelihood of the next token to generate for the option identifiers. However, this may not be the best setting to evaluate LLMs. Wang et al. [9] studied this on Instruction-tuned LLMs by training a classifier to predict which possible option to associate with the generated answer. This was done to glance over additional text generated by the model (e.g. the generated text could be "The answer is B." as opposed to the simple "B." token). They found that the log-likelihood and the generated text decisions were often not matching.

Regarding Italian evaluation, some works have approached this challenge. Bacciu et al. [10] released another version of the *Open Italian LLM Leaderboard*, considering a different variety of tasks. Mercorio et al. [11] released a benchmark based on questions that can be found in the INVALSI test, an Italian educational test, to further test the knowledge and reasoning abilities of these models on a dataset that is natively in Italian rather than obtained through machine translation. The latter is one of the main problems when evaluating these models, due to the lack of resources w.r.t. English language, datasets that are used at the state-of-the-art are translated using machine translation models. Still, all this effort made to evaluate Italian-adapted LLMs mainly relies on *closed-ended* tasks.

3. Experiments

We study pre-trained and language-adapted models to test their capabilities in the resolution of Italian language tasks. Specifically, we want to modify the typical formatting that is used in *multiple choice question answering* to study if the models are capable of correctly following and generating Italian text. Usually, the format shown in Listing 1 is used, where <QUESTION> is the question the model has to answer, <IDENTIFIER_i> and <OPTION_i> are the option identifier, which is usually a letter or a number, and the text of the possible answer to the previously provided question respectively. <CORRECT_IDENTIFIER> is the identifier of the option that is the correct answer to the question.

```
<QUESTION>:
<IDENTIFIER_1> <OPTION_1>
<IDENTIFIER_2> <OPTION_2>
...
<IDENTIFIER_N> <OPTION_N>

<CORRECT_IDENTIFIER>
```

Listing 1: closed-ended format

We aim to modify the task so that the model has to generate the text of the correct option instead of the identifier. To do so, we consider two main evaluation settings:

- **Open-ended** (OE): we remove the available options and only supply the question in the prompt;
- **Closed-ended no identifiers** (CE-NI): we format the options without an identifier, the model has to write the corresponding text of the correct option.

In particular, for the CE-NI setting, we apply the format shown in Listing 2, where <CORRECT_OPTION> is the text of the option that represents the correct answer to the question.

```
<QUESTION>:
<OPTION_1>
<OPTION_2>
...
<OPTION_N>

<CORRECT_OPTION>
```

Listing 2: closed-ended no identifiers format

<CORRECT_IDENTIFIER> and <CORRECT_OPTION> are the outputs that we expect the evaluated model should generate.
We provide complete examples of the prompt formats in Appendix A.

Generally models are also evaluated by calculating the log-likelihood rather than generating text directly. The chosen option is then selected based on the highest value. We choose to perform a generative task instead, to check whether the models are capable of generating the answer string only without additional text and to also check if they generate something outside of the provided options. To evaluate this case, we use the BLEU, ROUGE-L and BertScore F1 metrics, which are reference metrics used to evaluate the correspondence of a generated sentence with a base one. BLEU and ROUGE-L focus on matching n-grams, while BertScore leverages pre-trained Bert models to assess the semantic similarity between words of the two texts. Furthermore, we consider four different possible prompt formats:

- **Plain** (P): there is no formatting, the text of the task is provided as it is in the prompt, only a "Risposta:" string is added at the end;
- **Plain few-shot** (P-F): same as P, but multiple examples of input-output are provided;
- **Instruct** (I): the chat template of the model is applied to the text of the task;
- **Instruct few-shot** (I-F): same as I, but multiple examples of input-output are provided.

Furthermore, for the few-shot formats, we consider two distinct numbers of examples to provide in the prompt: one-shot and five-shots. The intuition is that a language-adapted LLM should significantly improve performance even when provided with a single example.

We consider these prompt formats because most of the evaluation settings for Italian LLMs are done without applying the chat template. We argue that this choice may not be the best one when considering *Instruct* models that have been trained using a specific prompt format to continue a conversation. They should be evaluated using the same prompt format since it is also the one that will be used in case of deployment.

To set up the experimental protocol, we use the *lm-evaluation-harness* library [6], which provides an immediate and intuitive command line to automatically evaluate LLMs on previously defined as well as custom tasks. Specifically, we define custom tasks within the library following the previously defined evaluation settings. To do so, we consider the following datasets:

- **ARC-Challenge** [12]: consists of multiple-choice science exam questions, the Challenge set consists of complex questions that were not correctly answered by both a retrieval and co-occurrence method;
- **MMLU** [13]: consists of multiple-choice questions from 57 different topics (e.g. mathematics, computer science, and so on), requiring problem-solving abilities and knowledge to answer correctly;
- **EXAMS** [14]: consists of multiple-choice questions from high school exams. The dataset contains different subsets curated for different languages and optionally contains additional paragraphs regarding the question (extracted from Wikipedia);
- **WWBM** [15]: consists of multiple-choice questions spanning a wide range of topics. The questions come from the Italian version of the *"Who Wants to Be a Millionaire?"* board game where contestants answer progressively difficult questions. The question-answer instances are split into different categories depending on the difficulty of the question itself.

For the Italian version of these datasets, both EXAMS and WWBM are provided with splits in the Italian language natively. For ARC and MMLU, instead, we use the Italian version provided in the library for the *okapi* task released by Lai et al. [16], who performed automatic translation of the original datasets using *GPT-3.5 Turbo* for several languages. For all of these datasets, we define two custom tasks which apply the OE and CE-NI evaluation settings automatically. The examples used in the few-shot settings are taken from the validation splits of the datasets. For EXAMS, we use the train split as a test split (since it is not provided), while for WWBM, we remove the first five instances from the original dataset and use them as a validation split.

Regarding the models, we experiment using the following:

- **Italia-9B-Instruct-v0.1**[2]: trained from scratch with a focus on the Italian language (90% of data in Italian and the rest in English) with instruction-tuning for conversational purposes;
- **LLaMAntino-2-chat-13b-hf-UltraChat-ITA** [17]: instruction-tuning of *LLaMAntino-2-chat-13b-hf-ITA* (an Italian-adapted LLM) using a translated version of the *UltraChat* dataset;
- **LLaMAntino-3-ANITA-8B-Inst-DPO-ITA** [18]: fine-tuning, DPO and adaptation using a mixture of Italian and English datasets starting from the *LLaMA-3-8B-Instruct* model;
- **maestrale-chat-v0.4-alpha-sft**[3]: instruction-tuning for 2 epochs on a conversational dataset consisting of 1.7M instances, starting from an Italian-adapted version of *Mistral-7b*;

[2] https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1
[3] https://huggingface.co/mii-llm/maestrale-chat-v0.4-alpha-sft

Model	Format	ARC_IT			MMLU_IT			EXAMS			WBMM		
		BLEU	ROUGE-L	Bert-Score	BLEU	ROUGE-L	Bert-Score	BLEU	ROUGE-L	Bert-Score	BLEU	ROUGE-L	Bert-Score
Italia-9B-Instruct-v0.1	P	0.00	0.05	0.69	0.00	0.30	1.96	0.00	0.38	2.13	0.76	27.70	70.17
	P-F 1	2.17	13.43	68.88	1.35	8.72	54.52	1.28	13.25	66.87	2.58	33.47	77.29
	P-F 5	3.50	17.95	73.30	2.17	12.94	70.27	2.18	15.60	72.29	7.54	38.56	83.18
	I	0.52	7.17	64.30	0.75	6.91	63.13	0.50	6.57	63.11	0.24	7.65	63.36
	I-F 1	0.57	6.99	64.33	0.70	7.08	63.35	0.50	6.59	63.25	0.22	6.93	62.63
	I-F 5	0.70	8.00	65.35	0.84	7.95	64.45	0.56	7.04	63.52	0.30	10.16	64.77
LLaMAntino-2-chat-13b-hf-UltraChat-ITA	P	1.01	11.35	66.12	1.28	10.34	61.10	0.84	10.43	64.86	0.57	20.59	69.17
	P-F 1	1.99	15.47	71.38	0.99	8.87	62.97	1.42	14.39	69.41	3.35	33.64	81.18
	P-F 5	3.49	18.71	73.97	2.69	14.51	71.32	2.29	16.78	73.47	9.93	35.82	83.21
	I	0.80	7.50	64.34	0.87	6.94	63.27	0.50	6.25	62.87	0.24	8.51	64.03
	I-F 1	0.95	9.70	65.93	1.02	8.03	63.96	0.71	8.59	64.53	0.36	11.43	66.13
	I-F 5	1.61	14.15	70.09	0.87	6.94	66.40	1.06	12.57	68.70	2.42	32.73	70.10
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	P	0.88	10.18	65.71	0.95	10.08	65.39	0.66	10.45	65.01	0.23	15.25	67.05
	P-F 1	1.91	14.99	70.49	0.81	8.42	62.37	1.48	16.42	70.67	1.84	34.75	81.27
	P-F 5	1.41	15.24	69.40	0.75	10.59	65.00	1.40	17.74	72.63	2.94	35.32	82.36
	I	0.74	8.10	65.34	0.78	8.05	64.44	0.37	6.13	62.75	0.20	8.38	63.05
	I-F 1	1.14	11.41	68.83	0.72	9.21	63.29	0.77	14.69	68.03	0.36	11.43	76.91
	I-F 5	1.84	14.74	71.50	1.10	11.87	68.81	0.88	15.10	71.28	1.32	33.09	81.10
maestrale-chat-v0.4-alpha-sft	P	1.26	11.35	65.29	1.50	10.47	57.25	1.03	12.23	60.84	0.76	27.70	70.17
	P-F 1	3.43	19.45	73.16	1.49	12.14	65.56	2.86	22.53	73.09	6.75	46.26	84.60
	P-F 5	5.33	21.29	74.59	**3.40**	**17.99**	**72.53**	**4.48**	**23.45**	**75.77**	**20.66**	**50.50**	**87.08**
	I	0.88	8.38	64,61	0.99	8.15	63.65	0.77	11.05	65.53	0.47	19.98	69.34
	I-F 1	1.43	11.77	68.04	1.34	9.73	65.38	1.12	14.93	68.31	1.70	39.04	80.08
	I-F 5	2.34	16.27	71.37	1.91	15.11	69.33	2.47	20.83	74.12	2.86	45.05	84.10
Meta-Llama-3-8B	P	0.74	7.18	61.89	0.75	7.32	61.02	0.57	5.73	60.63	0.21	11.63	63.49
	P-F 1	3.35	18.57	73.58	1.31	10.21	63.81	2.99	21.10	72.85	9.06	40.66	83.82
	P-F 5	**5.59**	**21.53**	**74.85**	3.23	17.39	72.42	3.16	21.32	74.70	16.34	45.18	85.85
Meta-Llama-3-8B-Instruct	P	0.92	10.10	65.38	1,04	10.03	64.90	0.71	9.03	64.55	0.22	12.92	65.58
	P-F 1	2.56	17.29	72.06	1.11	8.85	62.76	1.83	18.00	70.81	3.99	37.27	82.28
	P-F 5	4.50	19.70	73.98	3.26	16.67	72.42	3.57	21.11	74.86	9.40	39.28	84.04
	I	0.50	6.07	64.00	0.72	6.19	63.24	0.41	5.15	62.25	0.21	6.69	62.07
	I-F 1	0.81	9.62	65.87	1.07	9.64	65.42	0.76	10.96	65.29	0.64	23.33	71.47
	I-F 5	2.46	17.44	72.09	2.35	15.41	71.01	0.88	15.10	73.84	5.96	39.86	83.87
Minerva-3B-base-v1.0	P	0.39	4.76	59.43	0.42	4.65	58.24	0.25	4.09	58.78	0.10	3.22	58.07
	P-F 1	0.76	9.75	67.01	0.58	5.90	60.49	0.38	5.57	60.98	2.22	27.03	78.51
	P-F 5	2.61	14.08	71.22	1.57	8.92	64.40	2.01	13.65	70.64	10.65	33.59	82.32
zefiro-7b-dpo-ITA	P	0.72	4.10	66.25	1.04	10.69	65.11	0.65	9.31	65.32	0.65	9.31	67.70
	P-F 1	3.64	16.47	72.60	1.19	11.31	66.58	2.75	17.09	71.21	6.12	33.15	81.85
	P-F 5	2.86	17.44	74.66	2.91	15.25	72.26	3.14	19.21	74.44	10.59	35.31	83.31
	I	0.65	6.96	63.50	0.85	6.91	62.85	0.55	6.23	62.47	0.22	6.96	63.20
	I-F 1	1.03	9.57	66.31	0.76	6.20	62.23	0.80	8.66	64.65	0.30	8.32	64.41
	I-F 5	1.91	14.50	70.63	1.91	15.11	66.09	1.52	15.36	70.47	0.81	24.60	73.30
LLaMA3-BILINGUAL *(Ours)*	P	0.80	9.17	64.41	1.00	9.34	64.13	0.67	8.32	63.68	0.20	11.77	64.80
	P-F 1	2.54	17.65	72.12	1.12	9.05	62.93	1.81	18.15	70.87	4.53	37.43	82.58
	P-F 5	4.69	19.68	74.09	3.26	16.89	72.24	3.31	20.85	74.61	9.54	39.35	84.03
	I	0.54	6.16	64.05	0.73	6.35	63.20	0.34	5.18	62.17	0.21	6.62	61.95
	I-F 1	0.90	10.63	66.72	1.19	10.48	65.88	0.91	12.63	66.24	0.77	27.20	73.93
	I-F 5	3.33	18.00	72.76	2.90	15.80	71.69	2.64	18.73	73.84	7.23	39.75	83.97
LLaMA3-ITA-ONLY *(Ours)*	P	0.87	6.75	64.07	0.97	9.10	64.59	0.64	7.78	63.23	0.19	10.51	64.02
	P-F 1	2.47	17.74	72.03	1.14	9.13	63.00	1.73	17.94	70.77	4.67	37.67	82.69
	P-F 5	2.61	16.64	74.10	3.11	16.97	72.21	3.22	21.04	74.65	8.91	39.34	84.05
	I	0.58	6.05	64.12	0.73	6.35	63.24	0.35	5.21	62.17	0.21	6.90	62.14
	I-F 1	1.02	10.94	67.03	1.26	10.79	66.33	0.96	12.95	66.52	0.77	27.20	74.25
	I-F 5	3.13	18.35	72.89	2.98	15.87	71.76	2.72	18.45	73.86	7.23	39.75	84.11

Table 1
Results for the OE setting. For the few-shots formats, the number of given shots is also provided next to the format name. The best result for each dataset and for each metric is in bold

- **Meta-Llama-3-8B**[4] and **Meta-Llama-3-8B-Instruct**[5]: latest version of the LLaMA family of models released by META (base and instruct version respectively);
- **Minerva-3B-base-v1.0**[6]: trained from scratch to be a proficient bilingual base model (English and Italian);
- **zefiro-7b-dpo-ITA**[7]: based on *zephyr* by Tunstall et al. [19], DPO training done on top of *zefiro-7b-sft-ITA*.

Furthermore, to test whether bilingual training helps the model solve these tasks, we instruction-tuned two new models. We start from the Meta-LLaMA-3-8B-Instruct checkpoint and fine-tune the model on $40,000$ instances from 3 different datasets: *databricks-dolly-15k*, *OpenOrca* and *UltraChat*. The datasets are automatically translated to Italian using ChatGPT 3.5. We consider two different settings, one where $20,000$ instances are kept for each language (Italian and English), and one where $40,000$ instances are kept for the Italian language only. For instruction tuning, we used LoRA with r equal to 16 and alpha equal to 16, targeting all linear layers of the model. Other hyperparameters are effective batch size equal to 128, learning rate equal to $2e-5$, weight decay equal to 0.01 and warmup steps equal to 5. In both cases, the instances to be used during the training are chosen at random.

For all experiments, we use the *greedy-decoding* generation strategy with a maximum number of tokens to

[4]https://huggingface.co/meta-llama/Meta-Llama-3-8B
[5]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[6]https://huggingface.co/sapienzanlp/Minerva-3B-base-v1.0
[7]https://huggingface.co/mii-community/zefiro-7b-dpo-ITA

Model	Format	ARC_IT			MMLU_IT			EXAMS			WBMM		
		BLEU	ROUGE-L	Bert-Score	BLEU	ROUGE-L	Bert-Score	BLEU	ROUGE-L	Bert-Score	BLEU	ROUGE-L	Bert-Score
Italia-9B-Instruct-v0.1	P	0.00	0.00	0.06	0.00	0.59	1.22	0.0	0.38	0.38	15.32	73.40	85.48
	P-F 1	53.48	55.09	87.09	36.80	49.17	84.18	55.49	55.00	86.74	45.60	55.00	82.55
	P-F 5	56.34	58.89	88.52	44.40	52.41	85.88	61.55	57.38	88.33	53.75	59.73	90.66
	I	5.76	21.91	71.17	9.00	27.68	72.64	4.32	18.44	68.91	0.80	20.14	69.70
	I-F 1	6.61	26.10	73.02	12.85	34.66	76.37	9.02	31.13	74.74	0.73	19.22	69.88
	I-F 5	20.48	42.83	81.79	17.92	40.90	80.14	28.41	47.58	83.99	13.18	48.74	87.45
LLaMAntino-2-chat-13b-hf-UltraChat-ITA	P	30.12	50.94	81.74	28.16	39.69	69.34	40.63	55.14	82.94	10.43	58.07	83.02
	P-F 1	55.05	61.92	86.97	31.61	49.91	82.15	55.25	61.98	85.13	63.84	68.91	90.84
	P-F 5	61.89	63.37	89.76	47.52	56.01	86.79	65.37	61.54	89.61	65.36	70.35	93.05
	I	12.48	28.34	72.03	9.86	20.21	68.39	7.87	22.46	69.09	1.24	22.45	69.34
	I-F 1	26.69	47.17	80.57	17.02	32.28	74.05	16.93	37.10	74.83	7.45	69.00	75.40
	I-F 5	45.81	57.95	86.78	30.61	48.57	82.92	42.04	51.42	82.78	36.48	65.88	91.00
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	P	12.15	37.28	74.72	14.69	37.91	75.05	12.21	38.12	75.46	1.30	39.35	76.48
	P-F 1	14.47	47.84	79.49	15.84	36.97	72.69	18.55	51.38	83.07	6.42	69.34	90.84
	P-F 5	22.85	61.81	85.17	15.85	47.98	79.34	17.64	56.84	84.49	7.37	68.90	91.11
	I	26.20	50.98	77.86	23.28	42.78	75.57	20.46	43.53	74.63	1.71	30.53	68.74
	I-F 1	20.74	55.60	84.26	15.74	40.51	75.90	17.07	49.49	81.87	3.89	63.97	88.29
	I-F 5	33.17	64.94	88.34	26.53	55.00	84.09	29.73	60.60	87.10	7.08	71.96	91.75
maestrale-chat-v0.4-alpha-sft	P	42.45	69.92	88.44	38.09	59.54	84.57	46.17	68.57	87.20	15.32	73.40	85.48
	P-F 1	79.53	79.04	94.04	34.92	55.74	83.36	62.81	71.17	87.53	69.73	78.49	94.88
	P-F 5	**81.20**	**80.55**	**94.59**	**62.02**	**68.65**	**90.72**	**72.63**	71.42	92.49	73.21	**79.76**	**95.18**
	I	16.11	34.10	73.41	12.34	24.07	69.21	7.91	28.05	70.58	2.52	32.78	73.04
	I-F 1	66.41	74.91	92.45	47.17	62.46	87.87	68.85	69.79	91.52	50.12	75.70	94.13
	I-F 5	78.44	77.93	93.85	59.44	67.17	90.14	71.50	70.67	92.14	71.27	77.23	94.60
Meta-Llama-3-8B	P	8.38	20.59	68.40	8.91	20.43	67.95	8.35	19.02	67.60	0.77	12.62	64.06
	P-F 1	70.20	72.06	92.15	26.07	48.25	80.63	67.09	66.66	90.67	70.29	73.23	93.71
	P-F 5	73.43	74.69	92.95	56.77	64.59	89.37	67.27	67.61	91.11	**73.73**	77.71	94.71
Meta-Llama-3-8B-Instruct	P	27.10	57.71	85.67	20.83	48.00	81.40	34.70	60.52	86.87	2.60	54.93	85.40
	P-F 1	69.96	74.04	92.17	22.95	41.62	75.98	57.83	65.96	85.58	65.54	74.66	94.09
	P-F 5	75.09	75.86	93.29	59.34	66.51	89.89	69.40	71.03	92.02	64.27	74.97	94.05
	I	27.30	46.34	87.41	17.68	29.85	70.09	14.68	35.41	71.00	2.97	36.10	68.84
	I-F 1	39.36	68.02	88.52	32.99	51.59	80.93	29.55	57.44	83.34	4.05	61.24	86.41
	I-F 5	76.67	77.67	93.89	61.79	67.93	90.33	70.09	**72.80**	**92.50**	31.83	78.24	94.61
Minerva-3B-base-v1.0	P	5.26	14.56	64.85	6.19	15.35	64.39	7.18	17.54	66.57	0.67	8.93	62.02
	P-F 1	24.75	38.08	81.24	15.42	31.38	76.28	35.85	42.49	83.13	26.74	38.71	85.39
	P-F 5	27.42	35.87	80.43	30.94	40.03	81.48	67.27	67.61	83.40	35.45	41.20	86.05
zefiro-7b-dpo-ITA	P	17.93	45.89	81.26	15.32	36.77	77.20	26.47	51.89	85.01	3.62	54.89	87.08
	P-F 1	62.63	67.49	89.74	46.24	55.33	86.50	57.02	61.54	85.34	56.91	65.59	91.97
	P-F 5	69.99	70.81	91.91	54.02	61.06	88.43	66.22	63.98	90.51	60.84	68.44	92.63
	I	4.95	15.47	66.80	5.47	14.85	65.80	6.04	16.51	66.77	1.40	43.83	65.65
	I-F 1	47.00	62.58	86.61	18.34	37.69	75.45	49.06	59.85	83.95	5.12	51.55	84.52
	I-F 5	61.73	68.53	89.21	59.44	67.17	86.33	55.84	64.23	87.26	5.70	58.93	87.96
LLaMA3-BILINGUAL *(Ours)*	P	14.41	43.85	79.53	14.00	38.01	76.92	20.49	52.95	83.29	1.40	43.83	80.01
	P-F 1	69.27	73.89	92.13	22.31	40.91	75.49	57.96	66.05	85.38	67.20	74.25	94.00
	P-F 5	73.31	75.04	93.08	59.53	66.61	89.95	69.32	70.60	91.93	65.09	74.98	94.07
	I	27.77	48.26	76.39	19.12	32.17	70.85	15.90	37.02	71.55	2.74	35.59	68.78
	I-F 1	40.94	69.83	89.47	34.58	54.21	82.18	37.44	62.63	86.22	6.78	68.31	90.47
	I-F 5	76.35	77.70	93.89	61.68	68.25	90.48	71.01	72.55	92.40	38.00	78.90	94.83
LLaMA3-ITA-ONLY *(Ours)*	P	12.60	38.93	77.42	13.08	35.94	75.97	17.48	49.55	81.90	1.22	39.87	78.14
	P-F 1	68.11	73.95	92.28	22.34	40.98	75.53	58.79	67.01	85.64	67.05	74.22	93.98
	P-F 5	73.05	75.14	93.07	59.40	66.68	89.96	69.87	70.98	92.02	67.14	75.68	94.26
	I	26.77	48.26	76.15	17.97	30.46	70.25	15.82	36.76	71.42	2.72	35.58	68.78
	I-F 1	45.48	71.08	89.89	37.10	55.43	82.88	43.47	64.79	87.24	7.45	68.99	90.73
	I-F 5	76.54	77.74	93.88	61.49	68.09	90.39	71.05	72.36	92.37	43.92	78.88	94.93

Table 2
Results for the CE-NI setting. For the few-shots formats, the number of given shots is also provided next to the format name. The best result for each dataset and for each metric is in bold

generate equal to 64. This limit was set for computational requirements and the value was chosen after studying the datasets to assess the number of tokens required for each answer. There was no combination of tokenizer and dataset which had a 95% percentile greater than 50 for token count, therefore we can safely set the previously defined boundary. We also set *torch.bfloat16* and use *flash-attention-2* [20] to speed up the generation process. Inference was always done with batch size set to 1 to maximize the quality of the generated text.

Furthermore, we consider changing the number of few-shots that are given in the prompt. Our assumption is that the models may learn to follow the patterns given in the examples, and therefore the Italian language generation may become more likely thanks to the additional information conveyed in the prompt. We aim to mitigate this potential bias by decreasing the number of shots. Thus, the number of shots for all settings using a few-shot strategy was set to either 1 or 5.

We report the results of the OE setting in Table 1 and of the CE-NI setting in Table 2 and comment them in the following section.

3.1. Hardware and Software Configuration

Our experimental setup consisted of a multi-node cluster provided by Fastweb SpA and equipped with Nvidia H100 GPUs for distributed training and evaluation. We used a suite of open-source libraries, including Transformers from Hugging Face [21], which provides seamless integration with PyTorch [22] and DeepSpeed [23], as well

as Unsloth[8] and TRL [24]. This software stack has been instrumental in efficiently handling large data sets and complex models.

This configuration allowed for parallelization of computations, significantly reducing training and evaluation time. DeepSpeed optimized memory usage and communication between nodes, allowing us to effortlessly scale evaluation processes across multiple model architectures.

The hardware-software combination ensured efficient, cost-effective, and reproducible experiments, which are critical for comparing multiple models and training new ones efficiently.

3.2. Findings and Additional Tests

Analyzing the results, it is clear that the OE strategy did not yield very satisfactory results for BLEU and ROUGE-L. We associate this with the difficulty of generating a response matching exactly the ground truth when the text that can be generated is not constrained in any way. To further support this point, we can see that the BertScore of some experiments yields good results, hinting that the semantics of the content that has been generated is similar to that of the ground truth.

Regarding the CE-NI strategy, the obtained results are much better for all metrics. Therefore providing the options in the input prompt greatly helped the model in limiting its generation to follow the provided options. Surprisingly, with respect to the Italian leaderboard where fine-tuned versions of the LLaMA 3 family were shown to have much better results, here the results are in line with the base models (or even worse in some cases). Furthermore, one of the best-performing models is *maestrale-chat-v0.4-alpha-sft*, which consistently outperforms the LLaMA 3 models in most cases.

For both settings the obtained results show that providing input-output examples in the prompt greatly enhances the results for all settings.

For both settings, primarily Instruct models were used. Upon analyzing the generated results, we observed instances where the model provided the correct result but appended an additional substring (e.g., the model began explaining the reasoning behind its response). To assess if this might have affected the result, we performed an additional test where we checked if the ground truth string was a substring of the generated output (after removing punctuation and trailing whitespaces as well as lowercasing the two strings). We report the complete results in Appendix C. Overall, some models show an improvement in performance, but the results still do not beat *maestrale-chat-v0.4-alpha-sft*.

We provide some generation examples in Appendix B.

4. Conclusions and Future Works

We have carried out a study on the effectiveness of evaluation of Italian-adapted LLMs on *closed-ended* tasks, multiple-choice question answering tasks specifically. We have experimented with two settings: an *open-ended* one and a *closed-ended* one without option identifiers. The results show better performance for the latter. Furthermore, they also show that, with respect to the *Open Italian LLM Leaderboard*, there are significant differences regarding model performance. We can conclude that the evaluation of Italian-adapted models should follow a more rigorous procedure which does not mainly rely on *closed-ended* tasks. We release the code that was used on GitHub[9]. In the future, we plan to further work on the topic and attempt to define best practices for the evaluation of these models.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.

[2] C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, T. Wolf, Open llm leaderboard v2, https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

[3] 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 33 (2020) 1877–1901.

[4] K. Wróbel, SpeakLeash Team, Cyfronet Team, Open pl llm leaderboard, https://huggingface.co/spaces/speakleash/open_pl_llm_leaderboard, 2024.

[5] C. Park, H. Kim, D. Kim, S. Cho, S. Kim, S. Lee, Y. Kim, H. Lee, Open ko-llm leaderboard: Evaluating large language models in korean with ko-h5 benchmark, in: ACL Main, 2024.

[6] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite,

[8] https://github.com/unslothai/unsloth

[9] https://github.com/swapUniba/Closed-ITA-LLM-Evaluation

B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2021. URL: https://doi.org/10.5281/zenodo.5371628. doi:10.5281/zenodo.5371628.

[7] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, W. Yin, Large language models for mathematical reasoning: Progresses and challenges, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 2024, pp. 225–237.

[8] K. Sun, Y. Xu, H. Zha, Y. Liu, X. L. Dong, Head-to-tail: How knowledgeable are large language models (llms)? aka will llms replace knowledge graphs?, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 311–325.

[9] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, B. Plank, "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models, 2024. URL: https://arxiv.org/abs/2402.14499. arXiv:2402.14499.

[10] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let's push Italian LLM research forward!, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), 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.

[11] F. Mercorio, M. Mezzanzanica, D. Potertì, A. Serino, A. Seveso, Disce aut deficere: Evaluating llms proficiency on the invalsi italian benchmark, 2024. URL: https://arxiv.org/abs/2406.17535. arXiv:2406.17535.

[12] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, arXiv:1803.05457v1 (2018).

[13] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, Proceedings of the International Conference on Learning Representations (ICLR) (2021).

[14] M. Hardalov, T. Mihaylov, D. Zlatkova, Y. Dinkov, I. Koychev, P. Nakov, EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 5427–5444. URL: https://aclanthology.org/2020.emnlp-main.438. doi:10.18653/v1/2020.emnlp-main.438.

[15] P. Molino, P. Lops, G. Semeraro, M. de Gemmis, P. Basile, Playing with knowledge: A virtual player for "who wants to be a millionaire?" that leverages question answering techniques, Artificial Intelligence 222 (2015) 157–181. URL: https://www.sciencedirect.com/science/article/pii/S0004370215000259. doi:https://doi.org/10.1016/j.artint.2015.02.003.

[16] V. Lai, C. Nguyen, N. Ngo, T. Nguyen, F. Dernoncourt, R. Rossi, T. Nguyen, Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2023, pp. 318–327.

[17] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, arXiv preprint arXiv:2312.09993 (2023).

[18] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, arXiv preprint arXiv:2405.07101 (2024).

[19] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, T. Wolf, Zephyr: Direct distillation of lm alignment, 2023. arXiv:2310.16944.

[20] T. Dao, FlashAttention-2: Faster attention with better parallelism and work partitioning, in: International Conference on Learning Representations (ICLR), 2024.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, 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.

[22] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. Luk, B. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan,

P. Wu, S. Chintala, Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation, in: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24), ACM, 2024. URL: https://pytorch.org/assets/pytorch2-2.pdf. doi:10.1145/3620665.3640366.

[23] C. Li, Z. Yao, X. Wu, M. Zhang, C. Holmes, C. Li, Y. He, Deepspeed data efficiency: Improving deep learning model quality and training efficiency via efficient data sampling and routing, 2024. URL: https://arxiv.org/abs/2212.03597. arXiv:2212.03597.

[24] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, Trl: Transformer reinforcement learning, https://github.com/huggingface/trl, 2020.

Appendix

A. Prompt Formats

All showcased examples in this section are obtained from Meta-Llama-3-8B-Instruct model.

Anna tiene un cubetto di ghiaccio. Perché si scioglie il cubetto di ghiaccio nella sua mano? Opzioni:
Il calore si sposta dalla sua mano al cubetto di ghiaccio.
Il freddo si sposta dalla sua mano al cubetto di ghiaccio.
Il calore si sposta dal cubetto di ghiaccio alla sua mano.
Il freddo si sposta dal cubetto di ghiaccio alla sua mano.
Risposta:

Example 1: Prompt in the P-F format for the OE setting

Le more selvatiche si riproducono asessualmente sprigionando nuove radici quando i loro steli toccano il terreno. Si riproducono anche sessualmente attraverso i loro fiori. Qual è il vantaggio della pianta di more di potersi riprodurre sessualmente e asessualmente? Opzioni:
Consente alle piante di crescere più in alto.
Produce fiori che attraggono gli insetti.
Produce more che hanno un sapore migliore.
Permette alle piante di more di adattarsi a nuove condizioni.
Risposta: Permette alle piante di more di adattarsi a nuove condizioni.

Anna tiene un cubetto di ghiaccio. Perché si scioglie il cubetto di ghiaccio nella sua mano? Opzioni:
Il calore si sposta dalla sua mano al cubetto di ghiaccio.
Il freddo si sposta dalla sua mano al cubetto di ghiaccio.
Il calore si sposta dal cubetto di ghiaccio alla sua mano.
Il freddo si sposta dal cubetto di ghiaccio alla sua mano.
Risposta:

Example 2: Prompt in the P-F 1 format for the OE setting

<|start_header_id|>user<|end_header_id|>

Anna tiene un cubetto di ghiaccio. Perché si scioglie il cubetto di ghiaccio nella sua mano? Opzioni:
Il calore si sposta dalla sua mano al cubetto di ghiaccio.
Il freddo si sposta dalla sua mano al cubetto di ghiaccio.
Il calore si sposta dal cubetto di ghiaccio alla sua mano.
Il freddo si sposta dal cubetto di ghiaccio alla sua mano.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Example 3: Prompt in the I-F format using LLaMA 3 chat template

<|begin_of_text|><|start_header_id|>user<|end_header_id|>

Le more selvatiche si riproducono asessualmente sprigionando nuove radici quando i loro steli toccano il terreno. Si riproducono anche sessualmente attraverso i loro fiori. Qual è il vantaggio della pianta di more di potersi riprodurre sessualmente e asessualmente? Opzioni:
Consente alle piante di crescere più in alto.
Produce fiori che attraggono gli insetti.
Produce more che hanno un sapore migliore.
Permette alle piante di more di adattarsi a nuove condizioni.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Permette alle piante di more di adattarsi a nuove condizioni.<|eot_id|><|start_header_id|>user<|end_header_id|>

Anna tiene un cubetto di ghiaccio. Perché si scioglie il cubetto di ghiaccio nella sua mano? Opzioni:
Il calore si sposta dalla sua mano al cubetto di ghiaccio.
Il freddo si sposta dalla sua mano al cubetto di ghiaccio.
Il calore si sposta dal cubetto di ghiaccio alla sua mano.
Il freddo si sposta dal cubetto di ghiaccio alla sua mano.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Example 4: Prompt in the I-F 1 format using LLaMA 3 chat template

B. Zero-shot Response Examples

All showcased examples in this section are obtained from Meta-Llama-3-8B-Instruct model.

Una sorgente sonora di frequenza f 0, si muove con velocità costante lungo una circonferenza. Nel centro della circonferenza si trova il ricevitore del suono. Quale asserzione è esatta per la frequenza f registrata dal ricevitore? Risposta:

Ground truth: f = f 0
Generated Answer: La frequenza f registrata dal ricevitore è costante e uguale a f 0

Example 5: Generated answer with additional text for the OE setting

Il periodo di rotazione di un satellite artificiale intorno ad un pianeta è T . La distanza tra il satellite ed il centro del pianeta è r . A che distanza dal centro del pianeta ruota un altro satellite se il suo periodo di rotazione è T / 8? Opzioni:
8 r
r/8
4 r
r/4
Risposta:

Ground truth: r/4
Generated Answer: r/8 Spiegazione: Se il periodo di rotazione del satellite è T / 8, allora la sua distanza dal centro del pianeta è r / 8. Infatti, se il satellite ruota in 8T, allora la sua distanza dal centro del pianeta è

Example 6: Generated answer with additional text for the CE-NI setting

C. Substring Matching Results

Model	Format	ARC_IT	MMLU_IT	EXAMS	WBMM
Italia-9B-Instruct-v0.1	P	0.00	0.26	0.20	45.47
	P-F 1	3.94	4.50	5.84	35.96
	P-F 5	5.73	5.00	5.84	36.78
	I	4.96	5.73	7.53	41.07
	I-F 1	4.53	5.86	7.72	41.38
	I-F 5	4.96	5.59	6.73	36.78
LLaMAntino-2-chat-13b-hf-UltraChat-ITA	P	6.07	5.91	7.13	32.69
	P-F 1	5.39	5.76	5.84	32.89
	P-F 5	5.82	5.88	7.03	32.12
	I	5.48	5.08	7.62	33.91
	I-F 1	5.90	6.28	7.23	34.48
	I-F 5	6.33	6.41	7.62	32.12
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	P	7.44	7.55	10.0	36.62
	P-F 1	7.10	6.58	8.42	34.02
	P-F 5	7.36	7.32	8.91	31.36
	I	4.96	5.89	7.82	36.42
	I-F 1	6.50	6.91	8.32	35.60
	I-F 5	6.07	6.66	6.63	30.90
maestrale-chat-v0.4-alpha-sft	P	7.02	7.49	10.69	45.47
	P-F 1	**8.30**	8.39	**11.68**	**47.16**
	P-F 5	8.13	8.53	11.58	45.01
	I	5.90	7.56	10.69	46.65
	I-F 1	7.19	8.00	10.59	46.29
	I-F 5	8.04	**8.60**	9.60	44.55
Meta-Llama-3-8B	P	5.48	6.95	9.11	37.85
	P-F 1	6.67	7.14	9.70	39.03
	P-F 5	5.73	7.35	9.70	40.0
Meta-Llama-3-8B-Instruct	P	7.96	7.65	10.0	38.26
	P-F 1	6.67	7.44	7.92	36.78
	P-F 5	6.76	7.54	10.0	35.35
	I	3.85	5.32	7.43	38.16
	I-F 1	6.16	6.07	9.80	40.56
	I-F 5	7.36	7.41	8.81	36.88
Minerva-3B-base-v1.0	P	2.57	3.48	4.46	30.49
	P-F 1	2.31	3.86	5.05	28.59
	P-F 5	3.34	2.74	4.36	30.54
zefiro-7b-dpo-ITA	P	5.39	6.20	2.18	29.67
	P-F 1	4.71	5.69	7.03	31.00
	P-F 5	4.96	6.56	8.42	31.56
	I	3.84	5.97	6.24	32.33
	I-F 1	5.82	4.98	6.83	28.54
	I-F 5	5.56	6.54	7.43	29.97
LLaMA3-BILINGUAL *(Ours)*	P	7.96	7.76	10.79	38.57
	P-F 1	6.84	7.54	8.12	36.68
	P-F 5	6.33	7.60	9.31	35.19
	I	3.85	5.47	7.82	38.47
	I-F 1	5.99	6.68	9.51	39.59
	I-F 5	7.36	7.50	8.22	36.57
LLaMA3-ITA-ONLY *(Ours)*	P	7.36	7.92	10.69	39.03
	P-F 1	7.02	7.57	8.02	36.78
	P-F 5	6.67	7.63	9.60	36.11
	I	3.94	5.48	7.82	38.21
	I-F 1	6.59	6.66	10.0	39.23
	I-F 5	7.36	7.59	7.62	36.47

Table
Sub-string matching results for the OE setting. For the few-shots formats, the number of given shots is also provided next to the format name. The best result for each dataset is in bold

Model	Format	ARC_IT	MMLU_IT	EXAMS	WBMM
Italia-9B-Instruct-v0.1	P	0.00	0.38	0.30	73.56
	P-F 1	39.86	33.19	37.53	52.43
	P-F 5	44.74	36.03	40.10	56.62
	I	29.77	29.59	26.73	55.91
	I-F 1	26.78	31.08	29.01	55.86
	I-F 5	32.59	31.42	32.77	56.62
LLaMAntino-2-chat-13b-hf-UltraChat-ITA	P	43.54	30.08	40.89	58.16
	P-F 1	49.10	38.17	44.65	66.19
	P-F 5	50.90	40.23	45.45	67.32
	I	41.66	26.29	34.75	60.56
	I-F 1	44.23	33.16	38.12	57.95
	I-F 5	48.08	39.50	36.83	62.92
LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	P	55.86	43.84	52.48	70.44
	P-F 1	60.57	45.34	48.32	72.38
	P-F 5	62.45	46.82	51.49	69.82
	I	61.85	44.93	54.46	75.91
	I-F 1	62.19	43.75	49.51	74.06
	I-F 5	61.42	45.11	52.87	75.14
maestrale-chat-v0.4-alpha-sft	P	69.38	50.18	58.71	73.56
	P-F 1	71.43	54.52	58.22	76.88
	P-F 5	**73.31**	**55.85**	58.02	**78.21**
	I	46.88	29.83	40.30	60.36
	I-F 1	69.63	52.22	56.54	74.58
	I-F 5	70.15	54.30	56.73	75.40
Meta-Llama-3-8B	P	57.57	46.30	56.54	75.09
	P-F 1	63.13	46.88	51.58	71.20
	P-F 5	66.47	50.49	53.37	75.96
Meta-Llama-3-8B-Instruct	P	59.54	44.26	53.07	68.85
	P-F 1	66.30	50.13	51.18	72.79
	P-F 5	68.69	52.42	57.43	72.79
	I	57.83	36.04	48.61	74.89
	I-F 1	69.29	48.14	54.46	75.40
	I-F 5	70.83	54.17	**60.10**	77.75
Minerva-3B-base-v1.0	P	47.48	43.71	59.90	73.86
	P-F 1	25.66	28.51	23.86	33.25
	P-F 5	20.10	23.09	22.87	34.94
zefiro-7b-dpo-ITA	P	48.76	39.18	41.58	60.67
	P-F 1	55.00	40.37	46.04	62.56
	P-F 5	60.31	45.34	48.42	64.86
	I	31.48	31.50	40.40	72.69
	I-F 1	50.98	46.11	45.15	66.55
	I-F 5	58.26	47.16	50.20	64.55
LLaMA3-BILINGUAL *(Ours)*	P	59.71	44.50	54.16	69.92
	P-F 1	66.04	49.70	50.89	72.53
	P-F 5	67.58	52.29	56.54	72.84
	I	60.65	38.61	50.20	75.35
	I-F 1	69.63	50.00	56.14	75.04
	I-F 5	70.49	54.51	**60.10**	77.90
LLaMA3-ITA-ONLY *(Ours)*	P	60.57	45.16	54.26	70.49
	P-F 1	66.21	49.79	51.98	72.43
	P-F 5	67.67	52.38	57.23	73.71
	I	59.88	37.08	50.40	75.40
	I-F 1	69.21	50.19	56.63	74.94
	I-F 5	70.40	54.28	59.41	77.65

Table
Sub-string matching results for the CE-NI setting. For the few-shots formats, the number of given shots is also provided next to the format name. The best result for each dataset is in bold