Evaluating Large Language Models on Italian Tasks

Bernardo Magnini[1,*], Roberto Zanoli[1,†], Michele Resta[2,†], Martin Cimmino[2,†], Paolo Albano[2,†], Marco Madeddu[3,†] and Viviana Patti[3,†]

[1]Fondazione Bruno Kessler, Italy
[2]Domyn, Italy
[3]University of Turin, Italy

Abstract

The rapid advancement of Large Language Models (LLMs) has highlighted the need for robust tools to evaluate them correctly. A major challenge in developing models that serve non-English speakers lies in the predominance of benchmarks that are either in English or machine translated from it. Evaluating the performance of multilingual or language-specific models requires native-language resources. In this paper, we present EVALITA-LLM a benchmark entirely composed of datasets in native Italian and adjusted to assess LLMs capabilities. The benchmark consists of 10 tasks that cover key aspects of NLP. We also provide prompts for all tasks that are designed to follow specific criteria. In order to avoid prompt sensibility, the evaluation of the models considers different methodologies to combine the scores obtained on different prompts.

Keywords

Benchmark, Italian, Evaluation, Large Language Models

1. Introduction

The creation and improvement of Large Language Models (LLMs) have greatly impacted the field of Natural Language Processing (NLP). The evaluation process for these models relies on having common benchmarks that can be applied to the different LLMs in order to compare them [1]. As of now, the benchmarks that are currently used concern different problems: general language understanding [2], mathematics [3] and more. A key feature of these datasets is that they are mostly in the English language, making them unfit to evaluate LLMs' abilities in less popular languages.

Recently, there have been different proposals to create LLMs for multiple languages [4] or for a single specific language (excluding English) [5]. This lead to the creation of a variety of models that also cover Italian, with more to come in the future.

To rigorously evaluate these models, we need benchmarks that are suited to both the Italian language and the generative nature of these models. Proposed benchmarks for the Italian language are the result of the application of Machine Translation (MT) to existing English datasets [6]. This approach presents different issues like poor quality of translations and the Anglo-American cultural contexts of the starting data. We also want to acknowledge the CALAMITA (Challenge the Abilities of LAnguage Models in ITAlian) initiative [7] which aims at creating a native Italian benchmark through the collaboration of the Italian research community.

Given these circumstances, we present Evalita-LLM, an Italian benchmark designed to evaluate generative models. The evaluation suite is composed of ten different tasks all in native Italian. The datasets we gathered were all already existing corpora that mostly come from the EVALITA, a recurring event to evaluate systems on Italian data, with a few exceptions. All tasks come with a set of prompts that

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
†These authors contributed equally.
magnini@fbk.eu (B. Magnini); zanoli@fbk.eu (R. Zanoli); michele.resta@domyn.com (M. Resta); martin.cimmino@domyn.com (M. Cimmino); paolo.albano@domyn.com (P. Albano); marco.madeddu@unito.it (M. Madeddu); viviana.patti@unito.it (V. Patti)
0000-0002-0740-5778 (B. Magnini); 0000-0003-0870-0872 (R. Zanoli); 0000-0002-7811-3895 (M. Resta); 0009-0004-5620-0631 (M. Madeddu); 0000-0001-5991-370X (V. Patti)

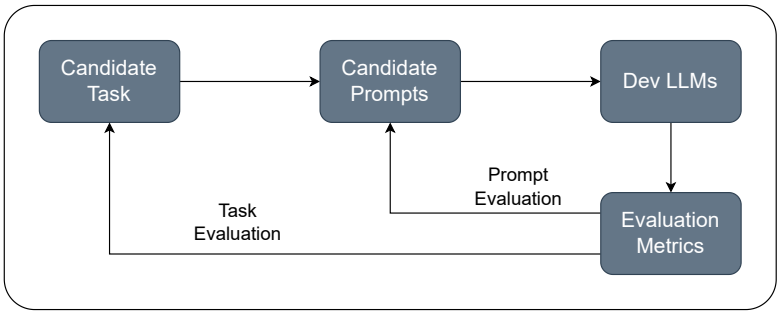

Figure 1: Evalita-LLM incremental validation methodology.

have been created by applying a specific methodology that we describe in Section 2.2. The benchmark has been entirely implemented in LM Evaluation Harness [8] and it is publicly available[1]. In addition, a leaderboard of recent LLMs with their performance on the Evalita-LLM benchmark is available on Hugging Face[2]. We hope this resource can become a reliable evaluation tool for Italian models.

2. Methodology

2.1. Task Selection

The entire EVALITA-LLM benchmarks consists of ten already existing native Italian datasets. Most of the corpora (eight) were part of an EVALITA challenge, with the remaining two being tasks that we evaluated as relevant enough to include in the benchmark. We include both *generative* and *multiple-choice* tasks. For the former, the model is prompted in a open generation setting, meaning that the output is a string of text that possibly needs post processing. The latter interrogates the model in a multiple-choice setting constraining the model to choose from a limited set of answers.

Given the large number of tasks that have been proposed in EVALITA throughout the years, we applied an incremental process (see Figure 1) to select the tasks to be included in the benchmark. The methodology works as follows:

- First, the candidate dataset (preferably available on the ELG catalog[3]), is converted into the format required by LM-Eval-Harness (JSON lines) and then it is uploaded to HuggingFace[4].
- Then, a task is defined on LM-Eval-Harness, including a pre-processing script, prompts creation, selection of examples for few-shot prompting, definition of the evaluation metrics and possible post-processing operations.
- Finally, the task was tested on a set of different LLMs. During this step we gathered data about the performance of the LLMs and their time of execution. The statistics were used to decide if a dataset or prompt was adequate to be included in the benchmark or not.

As for *multiple-choice*, at the end of the process, we selected six tasks: Word in Context (WIC), Textual Entailment (TE), Sentiment Analysis (SA), Hate Speech (HS), Frequently asked Questions (QA), and Admission Tests (AT). On the other side, we selected four *generative* tasks: Lexical Substitution (LS), Named Entity Recognition (NER), Relation Extraction (REL), and Summarization (SUM). Table 1 summarizes the main characteristics of the ten tasks, including the core linguistic competence that each task is supposed to assess, the domain of the task, whether it is multiple-choice or generative, and the metric used to calculate the score of the models.

[1] https://github.com/EleutherAI/lm-evaluation-harness
[2] https://huggingface.co/spaces/evalitahf/evalita_llm_leaderboard
[3] https://live.european-language-grid.eu/catalogue/
[4] https://huggingface.co/

#	Task	Core Competence	Domain	LLM Eval	Metric
1	Word in context	Word disambiguation	news	multiple-choice	F_1
2	Textual entailment	Semantic inference	news	multiple-choice	Accuracy
3	Sentiment analysis	Text classification	social	multiple-choice	F_1-macro
4	Hate speech	Text classification	social	multiple-choice	F_1-macro
5	FAQ	Question answering	P.A.	multiple-choice	Accuracy
6	Admission tests	Question answering	scientific	multiple-choice	Accuracy
7	Lexical substitution	Word disambiguation	news	generate-until	F_1
8	Entity recognition	Information extraction	mixed	generate-until	F_1
9	Relation extraction	Information extraction	scientific	generate-until	F_1
10	Summarization	Text generation	wiki	generate-until	Rouge

Table 1
Tasks in the Evalita-LLM benchmark.

2.2. Prompting Criteria

As generative LLMs can be sensitive to different prompting strategies [9], we established a set of general rules to design the prompts for the tasks, including that they are in Italian, they do not specify the role of the model, they are minimal and mention the type of input given to the model (e.g., a tweet). Given the above general rules, for each of the *multiple-choice* tasks we defined six different prompts. As an example, Table 2 shows the prompts for the Sentiment Analysis task.

	Pattern	Prompt	Options
p1	Question	What is the sentiment expressed in the following tweet:'{{text}}'?	[Positive, Negative, Neutral, Mixed]
p2	Task description + Question	You have to carry out a sentiment analysis task. What is the sentiment expressed in the following tweet:'{{text}}'?	[Positive, Negative, Neutral, Mixed]
p3	Question + Answer	What is the sentiment expressed in the following tweet:'{{text}}'? A: Positive \n B: Negative \n C: Neutral \n D: Mixed \n Answer:	[A, B, C, D]
p4	Task description + Question + Answer	You have to carry out a sentiment analysis task. What is the sentiment expressed in the following tweet:'{{text}}'? A: Positive \n B: Negative \n C: Neutral \n D: Mixed \n Answer:	[A, B, C, D]
p5	Affirmative	The following tweet: '{{text}}' expresses a sentiment that is	[Positive, Negative, Neutral, Mixed]
p6	Task description + Affirmative	You have to carry out a sentiment analysis task. The following tweet: '{{text}}' expresses a sentiment that is	[Positive, Negative, Neutral, Mixed]

Table 2
Prompts for the Sentiment Analysis task.

Meanwhile, we provided only two prompts for each *generative* tasks, selected among four patterns, reported in Table 3. This difference is due to the considerably longer times in interrogating LLMs in a generation setting.

	Pattern	Prompt
p7	Request	Summarize the following newspaper article: 'source' \n Summary:
p8	Task description + Request	You have to carry out an automatic synthesis task. Summarize the following newspaper article: 'source' \n Riassunto:
p9	Request + Output format	Extract all entities of type PER (person), LOC (place) e ORG (organization) from the following text. Report each entity with the following format: Entity$Type, separating each pair with ','. If there are no entities present, answer with '&&NOENT&&'. \n Text: '{{text}}' \n Entities:
p10	Task description + Request + Output format	You have to carry out a named entity recognition task. Extract all entities of type PER (person), LOC (place) e ORG (organization) from the following text. Report each entity with the following format: Entity$Type, separating each pair with ','. If there are no entities present, answer with '&&NOENT&&'. \n Text: '{{text}}' \n Entities:

Table 3
Generative prompts used for the SUM task (p7 and p8) and the NER task (p9 and p10).

3. Evaluation Metrics

In this Section we provide a brief description of the evaluation metrics used to score generative models in the Evalita-LLM benchmark. As the models are tested on multiple prompts, we employed different strategies of evaluation for each tasks:

- *Minimum performance*, which for a specific task assigns to the model the score of the prompt with the worst reported score.
- *Maximum performance*, which for a specific task assigns to the model the score of the prompt with the best reported score.
- *Average performance*, which for a specific task assigns to the model the score of the average calculated across all prompts.
- *Combined Performance Score* (CPS), which unites the maximum and average performance metrics. To define CPS, we first introduce a model saturation score:

$$Sat_M(M, T, IT) = 1 - (MaxP_M - AvgP_M) \tag{1}$$

$$Sat_I(IT, T, M) = 1 - (MaxP_I - AvgP_I) \tag{2}$$

This score measures how closely the model's best performance aligns with its average performance. A high saturation score indicates that the model's performance does not drop significantly for non-optimal instructions. Then, the CPS is calculated as the product of the model's best performance ($MaxP$) and its saturation (Sat):

$$CPS_M(M, T, IT) = Sat_M \cdot MaxP_M \tag{3}$$

Here, equation 4 represents the version of equation 3 for prompt combination:

$$CPS_I(IT, T, M) = Sat_I \cdot MaxP_I \tag{4}$$

4. The Evalita-LLM Benchmark

To develop the Evalita-LLM benchmark, we conduct several experiments using few currently available LLMs. We selected six LLMs as "dev" with similar characteristics: they are open source and available on Hugging Face, all of them are in range 7B-9B, they are instructed versions to ensure reasonable interpretation of prompt instructions, and they have been pre-trained on some Italian data. Since the goal of the experiments is to test and refine our methodological approach (i.e., task and prompt selection) rather than to rank the models by their performance, in the remainder of the paper, we keep the LLMs anonymized and refer to them using placeholders (i.e., LLM-1 to LLM-6).

	LLM-1	**LLM-2**	**LLM-3**	**LLM-4**	**LLM-5**	**LLM-6**	**MinP**	**MaxP**	**AvgP**	**CPS**
p1	55.00	68.25	45.25	64.50	75.50	64.00	**45.25**	75.50	62.08	65.37
p2	55.00	56.50	55.00	59.00	78.75	69.25	55.00	**78.75**	62.25	65.76
p3	70.25	64.75	55.00	49.25	73.25	60.50	49.25	73.25	62.17	65.13
p4	65.00	63.50	55.00	61.25	74.75	61.75	55.00	74.75	**63.54**	**66.37**
p5	55.75	54.25	55.00	60.50	57.50	49.00	49.00	60.50	55.33	57.37
p6	55.00	59.00	55.75	57.75	60.75	45.50	45.50	60.75	55.63	57.64
MaxP	70.25	68.25	55.75	64.50	**78.75**	69.25				
AvgP	59.33	61.04	53.50	58.71	**70.08**	58.33				
CPS	62.58	63.33	54.50	60.76	**71.93**	61.69				

Table 4
Results during the development phase: zero-shot F_1 on the Textual Entailment (TE) task.

As an example of multiple-choice task, Table 4 presents the results of the development phase for the Textual Entailment (TE) task. To validate the task, we consider the performance of six dev LLMs. As for baselines for the task, we consider the random guess (50.00). All of the six dev LLMs outperform the two baselines (Average performance over the six prompts ranges from 53.50 to 70.08), with the Maximum performance (MaxP) as 78.75 for LLM-5. These results confirm that the task is well understood by the LLMs used for dev, while still being challenging for them. On the prompt side (top-right in Table 4), we notice that accuracy scores range from a minimum of 45,25 for prompt *p1* to a maximum of 78.75 for prompt *p2* among the six models, while prompt *p4* has both the highest average (63.54) and the highest CPS (66.37). In addition, there is high variability of scores within the same LLM: for instance, LLM-1 scores 55.00 with *p1* and *p2*, and 70.25 with *p3*. Finally, the MaxP on single LLMs is achieved by five different prompts, out of the six we employed, showing that different models have different reactions to our prompts.

5. Conclusions

In this paper, we presented EVALITA-LLM, a new benchmark specifically designed for generative models and entirely consisting of Italian data. We created this resource starting from previous EVALITA tasks and applying changes in order to make them suited to evaluate generative models. In the end, we selected a total of ten tasks that have been fully implemented and ready to use on LM-Eval-Harness. The selected datasets try to cover the fundamental tasks of NLP and are differentiated in multiple-choice and generative tasks. In order to create a robust benchmark, we provide multiple prompts for each task in order to avoid having an evaluation that is prompt-sensitive. We also provided a range of evaluation metrics that try to mitigate the impact of prompt sensibility. The dataset as well as task definitions are all available online. In addition, a leaderboard for the benchmark is public[5], including, at the time of writing, more than 40 models tested on the benchmark.

[5] https://huggingface.co/spaces/evalitahf/evalita_llm_leaderboard

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

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] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, arXiv preprint arXiv:2009.03300 (2020).

[3] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, arXiv preprint arXiv:2103.03874 (2021).

[4] L. Qin, Q. Chen, Y. Zhou, Z. Chen, Y. Li, L. Liao, M. Li, W. Che, P. S. Yu, Multilingual large language model: A survey of resources, taxonomy and frontiers, arXiv preprint arXiv:2404.04925 (2024).

[5] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[6] L. Moroni, S. Conia, F. Martelli, R. Navigli, Towards a more comprehensive evaluation for Italian LLMs, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 584–599. URL: https://aclanthology.org/2024.clicit-1.67/.

[7] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the abilities of LAnguage models in ITAlian, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 1054–1063. URL: https://aclanthology.org/2024.clicit-1.116/.

[8] 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, The language model evaluation harness, 2024. URL: https://zenodo.org/records/12608602. doi:10.5281/zenodo.12608602.

[9] S. Anagnostidis, J. Bulian, How susceptible are llms to influence in prompts?, arXiv preprint arXiv:2408.11865 (2024).