A Leaderboard for Benchmarking LLMs on Italian

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

[1]Fondazione Bruno Kessler (FBK), Via Sommarive 18, 38123 Povo, Trento, Italy
[2]Domyn, Via Principe Amedeo, 5, 20124 Milano, Italy
[3]University of Torino, Computer Science Department, Corso Svizzera 185, 10149 Torino, Italy

Abstract

We present Evalita-LLM, a comprehensive benchmark and leaderboard designed to evaluate Large Language Models (LLMs) on Italian tasks. Evalita-LLM covers ten native Italian tasks, including both multiple-choice and generative formats, and enables fair and transparent comparisons by using multiple prompts per task, addressing LLMs' sensitivity to prompt phrasing. The leaderboard supports both zero-shot and few-shot evaluation settings and currently reports results for 23 open-source models. Our findings show consistent performance improvements with few-shot prompting and larger model sizes. Additionally, more recent versions of LLMs generally outperform their predecessors. However, no single model excels across all tasks, which highlights the task-dependent nature of LLM performance. Notably, generative tasks remain significantly more challenging than multiple-choice ones. Hosted on Hugging Face, the Evalita-LLM leaderboard offers a public and continuously updated platform for benchmarking and transparent evaluation of LLMs.

Keywords

LLMs, Benchmarking, Leaderboard

1. Introduction

Leaderboards have become essential tools for assessing performance in the rapidly evolving landscape of Large Language Models (LLMs), offering standardized comparisons across a large variety of tasks, such as language understanding, dialogue, reasoning and code generation. Among available leaderboards, the Hugging Face Open LLM Leaderboard [1] is a popular and widely used resource for researchers, particularly in the open-source community. Now in its second version, it introduces more challenging and reliable benchmarks, including MMLU-Pro, GPQA, MuSR, MATH, IFEval, and BBH. Other notable platforms, such as Scale SEAL[2], Vellum.ai[3], and LLM-Stats.com[4], support evaluation efforts. In addition, open-source initiatives focused on human preference evaluation, like Chatbot Arena[5] and the Chatbot Arena LLM Leaderboard[6], are playing a key role in advancing the benchmarking landscape.

Although LLM benchmarks have driven significant progress, they currently show limitations that affect the fairness and completeness of the evaluations process. First, the focus on English, makes them less useful for testing models meant to serve other languages, including Italian. This is particularly relevant because of the recent growth of LLMs with a specific training on Italian, like for instance LLaMAntino [2], the Minerva family [3], Italia[7], Velvet[8] and the recent model MIIA[9]. On the other side, current benchmarks for Italian, as for instance Itabench[10], often rely on automatic translations of English datasets, which is non optimal, due to poor translation quality and cultural differences that make fair testing harder.

A second issue in benchmarking LLMs is that most benchmarks are based on a single-prompt approach (i.e., one prompt is arbitrarily selected for each task). However, it is well known that LLMs are very sensitive to how prompts are phrased [4, 5, 6], and that even small changes in wording can lead to big differences in performance, making single-prompt evaluations less reliable and harder to compare. For example, IberBench [7], a benchmark designed for Iberian languages, employs a single-prompt evaluation methodology. While this simplifies the evaluation pipeline, the authors acknowledge that alternative prompts could lead to different performance outcomes.

Third, the vast majority of current benchmarks rely

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy [1]

magnini@fbk.eu (B. Magnini); marco.madeddu@unito.it (M. Madeddu); michele.resta@domyn.com (M. Resta); zanoli@fbk.eu (R. Zanoli); martin.cimmino@domyn.com (M. Cimmino); paolo.albano@domyn.com (P. Albano); viviana.patti@unito.it (V. Patti)

[1]https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/
[2]https://scale.com/leaderboard
[3]https://www.vellum.ai/llm-leaderboard
[4]https://llm-stats.com/
[5]https://openlm.ai/chatbot-arena/
[6]https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard
[7]https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1
[8]https://huggingface.co/Almawave/Velvet-14B
[9]https://huggingface.co/Fastweb/FastwebMIIA-7B
[10]https://huggingface.co/collections/sapienzanlp/ita-bench-italian-benchmarks-for-llms-66337ca59e6df7d7d4933896

almost exclusively on multiple-choice tasks, drastically limiting the capacity to test the generative abilities of LLMs, which have been mainly trained on open-text generation. Although multiple-choice format simplifies scoring, it often require artificial task reformulations that hide the model's natural ability to generate text. In contrast, generative tasks, although better reflect real-world applications, they pose challenges, including less reliable evaluation metrics and inconsistent output formatting.

To address the above mentioned issues, we introduce Evalita-LLM[11], a comprehensive benchmark with its associated leaderboard, specifically designed to evaluate LLMs on Italian tasks. The benchmark includes a diverse set of carefully validated tasks and uses multiple prompts per task to ensure more consistent and reliable evaluations. All tasks are originally written in Italian, avoiding issues related to translation quality or cultural mismatches. The benchmark combines both multiple-choice and generative tasks, offering a balanced and practical way to assess the full range of model abilities. Evalita-LLM is supported by a public leaderboard hosted on Hugging Face[12], which allows to conduct fair comparisons between models and tasks and helps the community to better understand how Italian LLMs perform and can be improved. The results on the Leaderboard confirm that using few-shot context-learning works better than using no examples (zero-shot) for most of the models. Results also confirm that bigger and newer models usually perform better, showing how fast LLMs are improving.

2. Benchmarking Methodology

The Evalita-LLM benchmark is created using existing datasets almost exclusively from the Evalita campaigns[13], supported by the Italian Association for Computational Linguistics (AILC[14]). Over the past 15 years, Evalita has produced approximately 70 datasets covering various language tasks. Around 35 of these are freely available through the European Language Grid (ELG)[15], thanks to the Evalita4ELG project [8] led by the University of Turin.

We selected 15 native Italian datasets: half for multiple-choice tasks and half for open-ended ones. For each task, we created approximately 20 prompt candidates, adapted from similar tasks (often in English) and refined through several rounds of testing. The prompts were tested on various Italian LLMs using fixed evaluation metrics. During this process, prompts that resulted in weaker performances across the various models were discarded, and overly difficult tasks were also excluded.

[11] https://github.com/EleutherAI/lm-evaluation-harness
[12] https://huggingface.co/spaces/evalitahf/evalita_llm_leaderboard
[13] https://www.evalita.it
[14] https://www.ai-lc.it
[15] https://live.european-language-grid.eu

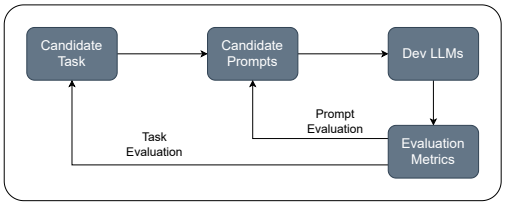

Figure 1: Evalita-LLM incremental validation methodology.

The Evalita-LLM benchmark was developed using the lm-evaluation-harness library[16] [9], which provides a unified interface for evaluating language models across a variety of tasks and formats. Since models' performance can be sensitive to their parameters, particularly *temperature* and *maximum context length*, the library allows users to adjust settings to some extent. In our setup, we follow the library's standard configuration to ensure consistency across evaluations. By default, temperature is set to `0.0`, resulting in deterministic (greedy) decoding, which favors reproducibility. To determine each model's input capacity, the maximum context length (the number of tokens a model can process per input) is retrieved dynamically by inspecting the model's configuration fields such as `n_positions`, `max_position_embeddings` or the tokenizer's `model_max_length`.

The benchmark construction followed three main steps:

- Dataset selection: datasets were converted into Hugging Face (HF) format and uploaded.
- Task definition: creating prompts, choosing few-shot or zero-shot, formatting output, and setting up metrics. The tasks are defined for evaluation only and are not used for model training.
- Model evaluation: tasks are tested on Italian LLMs during development to check if prompts work well.

Figure 1 shows how the benchmark was created step by step. At the end of the process, we selected ten tasks that cover different language types, text styles and real-world uses.

2.1. Prompting Approach

Prompt design is crucial since LLMs are highly sensitive to minor wording changes [10, 11, 12, 4, 5]. To address this issue, Evalita-LLM combines three main strategies: setting general rules for prompt design, using a compositional method to build prompts, and applying multiple prompts per task to ensure robustness and reliability.

[16] https://github.com/EleutherAI/lm-evaluation-harness

2.1.1. General Prompting Rules

The following rules guide the construction of prompts to ensure consistency, simplicity and alignment with the objectives of Evalita-LLM. The exact prompts used for each task are available on the leaderboard webpage[17]. Additional examples translated in English can be found in Appendix A.

- Prompts are entirely in Italian, including output labels.
- We avoid assigning roles to the model (e.g., "You are an assistant…").
- Prompts are short and simple to reduce bias.
- Each prompt specifies the type of input for the specific task (e.g., tweet, news, sentence).

2.1.2. Compositional Prompting

To ensure flexibility and systematic variation, we adopt a compositional approach, building each prompt from a combination of key elements:

- Core question or instruction (this is required for all prompts);
- High level task description (optional);
- Answer options (optional, for multiple-choice tasks);
- Output format instructions (optional, for generative tasks);

Keeping some components fixed reduces unnecessary prompt variations and simplifies evaluation. Around 20 templates were created for each task; after a testing phase (Section **??**), we kept 6 templates for multiple-choice and 4 for generative tasks, due to higher computational cost for generative evaluation.

2.1.3. Multiple Prompts for Multiple-choice Tasks

For multiple-choice tasks, we use six distinct prompt templates, each adapted to the specific task. The templates systematically vary the inclusion of a task description, the core question and the answer options:

- *Prompt 1: Question.* A base question that the model must answer, following general prompt guidelines.
- *Prompt 2: Task description + Question.* A brief task description is prepended to the question.
- *Prompt 3: Question + Answer.* The possible answers are appended to the question.
- *Prompt 4: Task description + Question + Answer.* This combines both the task description and the answer options with the question.
- *Prompt 5: Affirmative.* A simple affirmative statement that implicitly asks for an answer, without listing options.
- *Prompt 6: Task description + Affirmative.* The task description is prepended to the affirmative statement.

It has to be noted that in multiple-choice prompts, the answer options can be either explicitly embedded in the prompt or provided as options for evaluation process.

To minimize bias in model evaluation, attention was given to the order of answer choices in multiple-choice prompts. Only Prompt 3 and Prompt 4 are susceptible to such bias, as they explicitly list options (A, B, C, etc.). For tasks with fixed answer sets like Textual Entailment, options were kept in a natural order (e.g., A: True, B: False) to reflect typical human presentation. In contrast, for tasks with more open-ended answers, such as Admission Tests, the answer choices were shuffled during dataset creation to reduce positional bias.

2.1.4. Multiple Prompts for Generative Tasks

Generative prompts require the model to produce textual output, which is then evaluated for correctness using appropriate metrics. We adopt a compositional approach involving three key elements: (i) a mandatory request expressing the task; (ii) an optional brief task description placed at the beginning; (iii) optional output format instructions at the end.

Because generative tasks are computationally more expensive than multiple-choice tasks, we created four prompt types, which have been tested pairwise in our tasks. Tasks that need structured outputs get clear formatting instructions to help with parsing and scoring, while others allow freer text generation. The four prompt types are:

- *Prompt 7: Request.* A base generative request adhering to the general prompting guidelines.
- *Prompt 8: Task description + Request.* Adds a short task description before the request.
- *Prompt 9: Request + Output format.* Adds explicit instructions on the required output format.
- *Prompt 10: Task description + Request + Output format.* Combines the description, request, and output format instructions.

This modular design balances prompt diversity and evaluation efficiency across generative tasks.

2.1.5. Few-Shot Prompting

Few-shot prompting helps to improve performance by adding few examples of inputs and their corresponding correct responses within the prompt. For Evalita-LLM,

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

we used a 5-shot learning method. Except for Relation Extraction (REL) and Named Entity Recognition (NER), five examples were automatically selected from the training sets using LM-evaluation-harness. For REL and NER, examples were manually chosen to ensure full label coverage and output diversity, as many sentences for the two tasks do not contain any relevant entity or relation.

2.2. Evaluation Metrics

To select effective prompts for each task in Evalita-LLM, we adopt four prompt-scoring metrics inspired by [4]: *maximum*, *average*, *minimum*, and *combined performance*. These are used both to evaluate models over prompts and prompts over models.

Let M be an LLM, $T = \{(x_i, y_i)\}$ a task, I_T a set of prompts for T, and $\epsilon(M, T, i) \in [0, 1]$ the model's performance on task T with prompt i.

Minimum Performance Lowest performance of a prompt across all models:

$$MinP_I(I, T, M_T) = \min_{m \in M_T} \epsilon(I, T, m) \quad (1)$$

Maximum Performance Best performance of a model across prompts:

$$MaxP_M(M, T, I_T) = \max_{i \in I_T} \epsilon(M, T, i) \quad (2)$$

Best performance of a prompt across models:

$$MaxP_I(I, T, M_T) = \max_{m \in M_T} \epsilon(I, T, m) \quad (3)$$

Average Performance Mean model performance over prompts:

$$AvgP_M(M, T, I_T) = \frac{1}{|I_T|} \sum_{i \in I_T} \epsilon(M, T, i) \quad (4)$$

Mean prompt performance over models:

$$AvgP_I(I, T, M_T) = \frac{1}{|M_T|} \sum_{m \in M_T} \epsilon(I, T, m) \quad (5)$$

Combined Performance Score (CPS) This score integrates both stability (robustness) and best observed performance. First, saturation is defined as:

$$Sat_M(M, T, I_T) = 1 - (MaxP_M - AvgP_M) \quad (6)$$

$$Sat_I(I_T, T, M) = 1 - (MaxP_I - AvgP_I) \quad (7)$$

Then, CPS for models and prompts:

$$CPS_M(M, T, I_T) = Sat_M \cdot MaxP_M \quad (8)$$

$$CPS_I(I_T, T, M) = Sat_I \cdot MaxP_I \quad (9)$$

These metrics filter out unstable or poor-performing prompts and assist in choosing prompt sets that balance reliability and top performance across language models.

3. Benchmark Leaderboard

The Evalita-LLM leaderboard is a comprehensive platform that evaluates LLMs on 10 Italian-language tasks, both multiple-choice and generative. The leaderboard displays detailed metrics for each model and task, such as average performance over all prompts, best prompt performance and a combined score balancing accuracy and prompt consistency. Tasks span through multiple-choice questions, like Hate Speech and Sentiment Analysis, as well as generative requests, including Named Entity Recognition and Summarization. For each task, results are reported per prompt and combined for overall ranking. Users can filter and compare models by attributes like few-shot learning setup. Currently, the leaderboard presents evaluation results for 23 open source models in both zero-shot and few-shot settings, with new models being added as they become publicly available on the Hugging Face platform.

To optimize leaderboard management, models are indexed by their Hugging Face name. Only new, previously unlisted models are considered for evaluation, while revisions of already indexed models are skipped to save computational resources. Likewise, models are not re-evaluated on updated datasets ensuring resources are used for assessing new models.

3.1. Evalita-LLM Tasks

Word in Context (WiC). The Word in Context (WiC) task, proposed at Evalita 2023[18], focuses on word sense disambiguation in context. It consists of two sub-tasks: binary classification and ranking. For LLM evaluation, we focus on the binary classification task aimed at determining whether a target word *w* has the same meaning in two sentences, *s1* and *s2*. The best-performing system in the original challenge achieved an F_1-macro score of 85.00. In our experiments, the following dataset[19] was used.

Textual Entailment (TE). The Recognizing Textual Entailment (RTE) task was introduced at Evalita 2009[20]. It involves determining whether a hypothesis sentence is logically entailed by a given text sentence. The dataset consists of sentences sourced from Italian Wikipedia revision histories, labeled as entailed or not. The best model achieved 71% accuracy. We adapted this dataset[21] for our experiments.

[18] https://wic-ita.github.io/task
[19] https://huggingface.co/datasets/evalitahf/wic
[20] https://www.evalita.it/campaigns/evalita-2009/tasks/textual-entailment
[21] https://huggingface.co/datasets/evalitahf/textual_entailment

Sentiment Analysis (SA). The SENTIment POLarity Classification (SENTIPOLC) task was introduced at Evalita 2016[22]. It focuses on sentiment analysis of Italian tweets and includes three subtasks: polarity classification, subjectivity classification and irony detection. The best model achieved an F_1-macro score of 66.38. Our study concentrates on polarity classification, which categorizes each tweet's sentiment as positive, negative, neutral or mixed. We use this processed dataset[23].

Hate Speech (HS). The HaSpeeDe 2 challenge at Evalita 2020[24] focuses on detecting hateful content in Italian tweets and news headlines, targeting specific groups such as immigrants, Muslims, and Roma. Top-performing BERT-based models achieved an F_1-macro score of 80.88 on Twitter data and 77.44 on headlines. We use the adapted dataset[25], which combines both sources.

Frequently Asked Questions & Question Answering (FAQ). The QA4FAQ task, introduced at Evalita 2016[26], focuses on retrieving the most relevant FAQ entry given a user query. Systems must identify the closest matching question from a database of FAQs and return its answer. We transformed the dataset[27] into a multiple-choice format with four candidate answers per query.

Admission Tests (AT). The Admission Test task, introduced in [13], is not part of the Evalita campaign. It consists of answering multiple-choice questions from Italian medical specialty entrance exams (SSM), where each question has five options and only one correct answer. The questions cover a wide range of medical topics and often require complex reasoning beyond factual recall. We use this adapted dataset[28].

Lexical Substitution (LS). Task A of the Lexical Substitution challenge at Evalita 2009[29] focuses on identifying the most appropriate synonym for a target word given its context, without relying on predefined sense inventories. Systems are required to produce contextually relevant lemmas as substitutes. Evaluation is based on two metrics: *Best*, which scores the top candidate, and *Out-of-Ten (oot)*, which considers the top 10 suggestions. The best system achieved an F_1 score of 7.64 for *Best* and 38.82 for *oot*. In our experiments, we use the processed dataset[30], and follow the *oot* evaluation setting

Named Entity Recognition (NER). The Named Entity Recognition task at Evalita 2023[31] focuses on identifying and classifying person, organization, and location entities in Italian texts from multiple domains. The dataset, derived from the Kessler Italian Named-entities Dataset, includes documents from three sources: Wikinews, Literature, and Political Writings. The best model achieved an F_1-macro score of 88%. We use this processed dataset[32] in our experiments.

Relation Extraction (REL). The CLinkaRT task at Evalita 2023[33] addresses relation extraction in the clinical domain, focusing on linking laboratory results (RML) to their corresponding test events (EVENT) in Italian medical narratives[14]. Systems were evaluated using Precision, Recall, and F_1 score, with the best model achieving an F_1 of 62.99. We use the processed dataset[34], where entity pairs are restricted to occur within sentence boundaries.

Summarization (SUM). The summarization task, based on the Fanpage dataset [15], involves generating concise summaries of Italian news articles. The dataset includes news articles with titles, abstracts, and full texts across 9 categories. In the original study, mBART models achieved ROUGE-1: 38.91 and ROUGE-2: 21.38. For evaluation, we use a 10% subset of the original dataset[35], from which 100 samples were randomly selected for testing.

3.2. Models' Performance

Table 2 summarizes the performance of 23 models on two different testing conditions: few-shot (FS) and zero-shot (ZS). In the FS setting, models are given a few examples to guide their responses, while in ZS, they are asked to perform tasks without prior examples. Each model's performance was evaluated using the specific accuracy measure employed in the original task, and the results are combined into an average combined performance score (AvgCPS) across all tasks. The best performing model in the FS setting is gemma-3-27b-it, achieving an AvgCPS score of 57.42, while the lowest is Minerva-7B-base-v1.0 with 35.06. In ZS, scores range from 50.29 AvgCPS (gemma-3-27b-it) down to 30.23 (Volare).

Table 3 compares model accuracy on specific tasks against established reference scores, which come from

[22] https://www.evalita.it/campaigns/evalita-2016/tasks-challenge/sentipolc
[23] https://huggingface.co/datasets/evalitahf/sentiment_analysis
[24] http://www.di.unito.it/~tutreeb/haspeede-evalita20/index.html
[25] https://huggingface.co/datasets/evalitahf/hatespeech_detection
[26] https://www.evalita.it/campaigns/evalita-2016/tasks-challenge/qa4faq
[27] https://huggingface.co/datasets/evalitahf/faq
[28] https://huggingface.co/datasets/evalitahf/admission_test
[29] https://www.evalita.it/2009/tasks/lexical
[30] https://huggingface.co/datasets/evalitahf/lexical_substitution
[31] https://nermud.fbk.eu
[32] https://huggingface.co/datasets/evalitahf/entity_recognition
[33] https://e3c.fbk.eu/clinkart
[34] https://huggingface.co/datasets/evalitahf/relation_extraction
[35] https://huggingface.co/datasets/evalitahf/summarization-fp

Table 1
Tasks in the Evalita-LLM benchmark. Each task is categorized by its core competence, domain, evaluation type, and metric used.

#	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	PA	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 2
Model performance in few-shot (FS) and zero-shot (ZS) settings, reported in terms of Avg. Combined Performance Score (AvgCPS). Models are sorted in descending order by FS AvgCPS.

Model	FS	ZS
gemma-3-27b-it	57.42	50.29
Qwen2.5-14B-Instruct-1M	55.12	44.36
gemma-3-12b-it	54.32	47.35
gemma-2-9b-it	54.04	47.54
Qwen2.5-7B-Instruct	53.02	45.50
phi-4	52.24	38.37
Llama-3.1-SuperNova-Lite	52.11	43.06
granite-3.1-8b-instruct	51.70	37.26
Phi-3-medium-4k-instruct	51.22	42.09
Meta-Llama-3.1-8B-Instruct	50.37	40.23
Phi-3.5-mini-instruct	50.06	44.40
Llama-3-8b-Ita	49.41	41.02
LLaMAntino-3-ANITA-8B	49.39	42.14
maestrale-chat-v0.4-beta	49.37	41.04
aya-expanse-8b	49.30	40.25
Mistral-7B-Instruct-v0.3	47.31	41.56
gemma-3-4b-it	46.57	44.59
Llama-3-8B-4bit-UltraChat	45.33	36.28
Volare	44.13	30.23
occiglot-7b-it-en-instruct	44.09	38.00
Velvet-14B	43.09	39.48
Minerva-7B-instruct-v1.0	35.70	32.50
Minerva-7B-base-v1.0	35.06	32.36

the best systems in previous Evalita shared tasks or original task publications. It is important to note that these reference scores were obtained using supervised approaches. That is, models were trained on the corresponding task-specific training data. In contrast, the models evaluated in this study were tested in zero-shot or few-shot configurations, without using any of the training data to fine-tune or train the models on the specific tasks. Despite this difference in setup, the results show that some tasks benefit substantially from the advances in LLMs: for example, Textual Entailment (TE) accuracy improves by over 22%, and Sentiment Analysis (SA) by nearly 22%. On the other hand, some tasks remain challenging. Named Entity Recognition (NER) shows a large accuracy drop of more than 53%, and Relation Extraction (RE) decreases by over 18%.

Table 3
Comparison between reference accuracies from Evalita benchmark systems and the best result across all models and all prompt variants. The last column shows the percentage change in model accuracy compared to the reference accuracy.

#	Task	Ref. Accuracy	Model Accuracy	Delta (%)
1	WiC	85.00	72.47	-14.73
2	TE	71.00	86.75	+22.04
3	SA	66.38	80.80	+21.69
4	HS	80.88	77.77	-3.86
5	FAQ	–	99.50	–
6	AT	82.40	90.40	+9.71
7	LS	38.82	45.55	+17.29
8	NER	88.00	40.72	-53.68
9	RE	62.99	51.56	-18.15
10	SUM	38.91	34.88	-10.36

Figures 2 and 3 show two important trends about model size and in-context learning ability. First, the accuracy values tend to increase with model size, although the trend is not strictly linear. Second, models with 5 to 15 billion parameters benefit the most from few-shot prompting.

Table 4 reports the best-performing mid-size model (up to 15B parameters) for each task, considering the best score achieved across both zero-shot (ZS) and few-shot (FS) configurations.

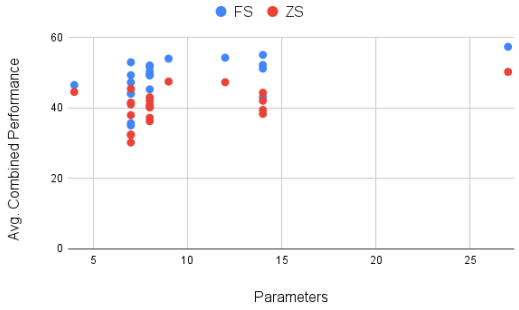

Figure 2: Comparison of model accuracy by size (in billions) and evaluation setting: zero-shot (ZS) vs. 5-few-shot (FS).

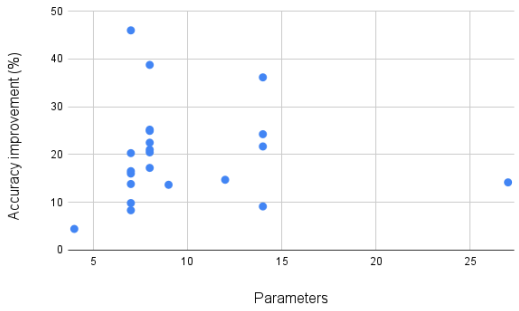

Figure 3: Accuracy gain (%) from zero-shot (ZS) to 5-shot (FS) evaluation versus model size.

Table 4
Best-performing mid-size model (<15B parameters) for each task, selected based on best score across zero-shot (ZS) and few-shot (FS) settings.

Task	Model
WIC	Phi-3-medium-4k-instruct
TE	Qwen2.5-14B-Instruct-1M
SA	gemma-3-12b-it
HS	Qwen2.5-14B-Instruct-1M
FAQ	LLaMAntino-3-ANITA-8B-Inst-DPO-ITA
AT	gemma-3-12b-it
LS	Lexora-Medium-7B
NER	Meta-Llama-3.1-8B-Instruct
REL	gemma-2-9b-it
SU	Velvet-14B

4. Discussion

In this section we analyze the results of the Evalita-LLM leaderboard across several perspectives to better understand the strengths and limitations of current LLMs on Italian tasks.

Zero-shot vs Few-shot Settings. Few-shot (FS) learning is examined from two complementary perspectives: the type of task and the size of the model. Figure 2 shows that models generally perform better in FS settings compared to zero-shot (ZS) ones. The gains are particularly significant in generative tasks, particularly Relation Extraction (RE) and Named Entity Recognition (NER), where the examples provided help the models to produce correctly formatted outputs. For example, in the RE task, *gemma-2-9b-it* (the best-performing model) improves its Combined Performance Score (CPS) from 34.97 in the ZS setting to 51.26 in FS. On the NER task, *Meta-Llama-3.1-8B-Instruct* increases its CPS from 7.93 to 40.3. In parallel, Figure 3 explores the relationship between model size and the accuracy gain from the ZS to FS setting. The most important improvements are observed in mid-sized models (approximately 5–15B parameters), which seem to benefit most from examples without being overly optimized, as may be the case with the largest models.

Model Size vs. Performance. Figure 2 shows a moderate positive correlation between the number of model parameters and accuracy. Specifically, the Pearson correlation coefficient is 0.4816 for the 5-shot setting and 0.4567 for the zero-shot setting. While larger models generally tend to achieve higher accuracy, the relationship is not strongly linear. This indicates that factors beyond model size, such as the model architecture, the quality of the training data and of the instruction tuning, significantly influence performance.

Performance Evolution within a Model Family. We compared two large language models from the same family, *Gemma-2 27B* and *Gemma-3 27B*, in both ZS and FS configurations. Our goal was to see whether performance improves from one generation to the next and to identify which tasks benefit most from the newer model. In the FS setting, *Gemma-3* shows the best overall performance, with the highest average CPS (57.42), which is 3.56 points higher than *Gemma-2*. In the ZS setting, however, *Gemma-2* slightly outperforms *Gemma-3* (50.60 vs. 49.89). Looking at individual tasks, *Gemma-3* performs better than *Gemma-2* in 9 out of 10 tasks in the FS setting, especially in: Relation Extraction (+11.9), Lexical Substitution (+7.6) and Sentiment Analysis (+6.0). In the ZS configuration, *Gemma-3* performs better on 6 out of 10 tasks, particularly in: Lexical Substitution (+6.37) and Hate Speech Detection (+4.88). *Gemma-2* outperforms *Gemma-3* on 4 tasks. Notably, Relation Extraction and Word in Context shows the largest gap in favor of *Gemma-2* (+34.8, +15, respectively). This result suggests that *Gemma-3* can be better effectively optimized for in-context learning and prompt-based fine-tuning.

Generative vs. Multiple-Choice Tasks. Generative tasks appear to be more challenging for large language models compared to multiple-choice tasks. Unlike multiple-choice format, where the output space is constrained and the model only needs to select among predefined options, generative tasks require models not only to understand the content of the request, but also to produce structured outputs in specific formats, which has then to be correctly parsed by a scoring script. As an example, formatting constraints in the Named Entity Recognition (NER) generative task poses significant challenges for LLMs, regardless of their ability to detect entities. When asked to output entities in the format $entity\$type$, models often fail in the zero-shot setting, with low output rates and formatting errors (e.g., using commas instead of the dollar sign as separator). Models improved performance with 5-shot prompting, mainly due to better adherence to the required output structure.

Additionally, evaluating generative outputs is difficult due to limitations in current metrics like BLEU and ROUGE, which focus on surface-level text overlap. Although advanced metrics like BERTScore and COMET consider context and meaning, they still cannot fully replicate human judgment. Combining multiple metrics might effectively mitigate these limitations by providing a more comprehensive assessment of task complexity from different perspectives.

To better understand how much harder generative tasks are for models, we compared their performance to reference scores from the Evalita benchmarking initiative (or the original dataset authors when Evalita scores were unavailable). Results in Table 3 confirm that while models often outperform reference baselines in multiple-choice tasks such as Textual Entailment (+22.04%), Sentiment Analysis (+21.69%), they have some difficulties in performing on generative tasks. For instance, model accuracy falls short in Named Entity Recognition (−53.68%) and Relation Extraction (−18.15). It is important to note, however, that the reference baselines were obtained using supervised models trained on task-specific datasets, whereas the models evaluated in this study were tested in zero-shot or few-shot settings, without any task-specific fine-tuning. These results further demonstrate how effectively modern LLMs can generalize to new tasks.

Model Specialization by Task. The results presented in Table 4 show that different models are better at different tasks. In fact, no single model achieves the best performance in all tasks, which means that performance crucially depends on the characteristics of the individual task. For example, *Qwen2.5-14B-Instruct-1M* performs as the best model on multiple-choice tasks as Textual Entailment and Hate Speech Detection, while *gemma-3-12b-it* performs best on Sentiment Analysis and the Admission Test.

5. Conclusion

This study introduced Evalita-LLM, a comprehensive benchmark and leaderboard designed to evaluate LLMs on Italian language tasks. The benchmarks and the evaluation metrics consider critical aspects of generative models (e.g., multiple-prompting, generative tasks output postprocessing,...).

Our findings show that few-shot settings generally outperform zero-shot settings, especially in generative tasks. This advantage is particularly noticeable in tasks such as Relation Extraction and Named Entity Recognition, where concrete examples help models produce correctly formatted outputs. We also found that mid-sized models benefit the most from few-shot learning. While there is a positive correlation between model size and accuracy, factors such as training data quality, and instruction tuning play significant roles. Additionally, newer versions within the same model family tend to outperform their predecessors on many tasks, but not all.

The publicly available Evalita-LLM leaderboard on Hugging Face can be used as a valuable resource for ongoing benchmarking and transparent comparison of emerging models on Italian tasks. The overall goal is to provide an evaluation tool that is easy to access and that can provide a fair assessment of a model and track difference in performance caused by different variables (model's size, model's version and more).

Limitations The number of datasets included for each task of the Evalita-LLM benchmark is limited in order to allow reasonable running times. In fact, the goal is not to create a repository that gathers all Italian datasets but rather to provide a tool for strong evaluation of models.

The metrics used for each tasks are the ones proposed in the original challenges and papers to allow for a direct comparison between systems. For this reason, we opted to not include more recent metrics such as BERT-score, which can be useful additions in the future.

Acknowledgments

This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by NextGenerationEU. The work of Marco Madeddu and Viviana Patti is partially supported by "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme. We warmly thank Alessandro Ercolani and Samuele Colombo for their invaluable support and guidance in writing the code and implementing this leaderboard.

References

[1] C. Bosco, E. Ježek, M. Polignano, M. Sanguinetti, Preface to the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[2] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. URL: https://arxiv.org/abs/2405.07101. arXiv:2405.07101.

[3] R. Orlando, L. Moroni, P.-L. H. 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: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 707–719.

[4] M. Mizrahi, G. Kaplan, D. Malkin, R. Dror, D. Shahaf, G. Stanovsky, State of what art? a call for multi-prompt llm evaluation, 2024. URL: https://arxiv.org/abs/2401.00595. arXiv:2401.00595.

[5] F. M. Polo, R. Xu, L. Weber, M. Silva, O. Bhardwaj, L. Choshen, A. F. M. de Oliveira, Y. Sun, M. Yurochkin, Efficient multi-prompt evaluation of llms, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 57 Morehouse Ln, Red Hook, NY 12571, United States, 2024, pp. 22483–22512.

[6] M. Sclar, Y. Choi, Y. Tsvetkov, A. Suhr, Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024. URL: https://arxiv.org/abs/2310.11324. arXiv:2310.11324.

[7] J. Ángel González, I. B. Obrador, Álvaro Romo Herrero, A. M. Sarvazyan, M. Chinea-Ríos, A. Basile, M. Franco-Salvador, Iberbench: Llm evaluation on iberian languages, 2025. URL: https://arxiv.org/abs/2504.16921. arXiv:2504.16921.

[8] V. Basile, C. Bosco, M. Fell, V. Patti, R. Varvara, Italian NLP for everyone: Resources and models from EVALITA to the European language grid, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 174–180. URL: https://aclanthology.org/2022.lrec-1.19/.

[9] S. Biderman, H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, A. DiPofi, J. Etxaniz, B. Fattori, J. Z. Forde, C. Foster, J. Hsu, M. Jaiswal, W. Y. Lee, H. Li, C. Lovering, N. Muennighoff, E. Pavlick, J. Phang, A. Skowron, S. Tan, X. Tang, K. A. Wang, G. I. Winata, F. Yvon, A. Zou, Lessons from the trenches on reproducible evaluation of language models, 2024. URL: https://arxiv.org/abs/2405.14782. arXiv:2405.14782.

[10] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3560815. doi:10.1145/3560815.

[11] G. Qin, J. Eisner, Learning how to ask: Querying LMs with mixtures of soft prompts, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 5203–5212. URL: https://aclanthology.org/2021.naacl-main.410. doi:10.18653/v1/2021.naacl-main.410.

[12] S. Sane, A. McLean, A notso simple way to beat simple bench, 2024. URL: https://arxiv.org/abs/2412.12173. arXiv:2412.12173.

[13] S. Casola, T. Labruna, A. Lavelli, B. Magnini, et al., Testing chatgpt for stability and reasoning: A case study using italian medical specialty tests., in: CLiC-it, 2023.

[14] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, Clinkart at EVALITA 2023: Overview of the task on linking a lab result to its test event in the clinical domain, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, Parma, Italy, 2023. URL: https://ceur-ws.org/Vol-3473/paper43.pdf.

[15] N. Landro, I. Gallo, R. La Grassa, E. Federici, Two new datasets for italian-language abstractive text summarization, Information 13 (2022). URL: https://www.mdpi.com/2078-2489/13/5/228. doi:10.3390/info13050228.

Table 5
Prompt patterns for the Sentiment Analysis task. Each prompt follows a different structure to test model robustness.

ID	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 6
Generative prompts used for the Summarization task (p7, p8) and the Named Entity Recognition task (p9, p10).

ID	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 Summary:
p9	Request + Output format	Extract all entities of type PER (person), LOC (place), and ORG (organization) from the following text. Report each entity in the format: Entity$Type, separated by ',. If there are no entities, respond with '&&NOENT&&'. \n Text: 'text' \n Entities:
p10	Task description + Request + Output format	You have to carry out a named entity recognition task. Extract all entities of type PER (person), LOC (place), and ORG (organization) from the following text. Report each entity in the format: Entity$Type, separated by ',. If there are no entities, respond with '&&NOENT&&'. \n Text: 'text' \n Entities:

A. Prompt Examples for Evalita-LLM Tasks

Table 5 presents different prompt structures for the Sentiment Analysis task, used here as an example of a multiple-choice task. Table 6 shows generative prompts for tasks such as Summarization and Named Entity Recognition.