

CALAMITA: Challenge the Abilities of LAnguage Models in ITALian

Giuseppe Attanasio^{1,*}, Pierpaolo Basile^{2,*}, Federico Borazio³, Danilo Croce^{3,*},
Maria Francis^{4,5}, Jacopo Gili⁶, Elio Musacchio², Malvina Nissim^{4,*}, Viviana Patti^{6,*},
Matteo Rinaldi⁶ and Daniel Scalena^{7,4}

¹Instituto de Telecomunicações, Lisbon, Portugal

²University of Bari “Aldo Moro”, Bari, Italy

³University of Rome “Tor Vergata”, Rome, Italy

⁴CLCG, University of Groningen, Groningen, The Netherlands

⁵University of Trento, Trento, Italy

⁶Computer Science Department, University of Turin, Turin, Italy

⁷University of Milan Bicocca, Milan, Italy

Abstract

The rapid development of Large Language Models (LLMs) has called for robust benchmarks to assess their abilities, track progress, and compare iterations. While existing benchmarks provide extensive evaluations across diverse tasks, they predominantly focus on English, leaving other languages underserved. For Italian, the EVALITA campaigns have provided a long-standing tradition of classification-focused shared tasks. However, their scope does not fully align with the nuanced evaluation required for modern LLMs. To address this gap, we introduce “Challenge the Abilities of LAnguage Models in ITALian” (CALAMITA), a collaborative effort to create a dynamic and growing benchmark tailored to Italian. CALAMITA emphasizes diversity in task design to test a wide range of LLM capabilities through resources natively developed in Italian by the community. This initiative includes a shared platform, live leaderboard, and centralized evaluation framework. This paper outlines the collaborative process, initial challenges, and evaluation framework of CALAMITA.

Keywords

Italian Benchmark, Shared Task, Language Models

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 – 06, 2024, Pisa, Italy

*Corresponding authors.

[†]These authors contributed equally.

✉ giuseppe.attanasio@lx.it.pt (G. Attanasio);
pierpaolo.basile@uniba.it (P. Basile); borazio@ing.uniroma2.it (F. Borazio); croce@info.uniroma2.it (D. Croce);
maria.francis287@gmail.com (M. Francis);
jacopo.gili584@edu.unito.it (J. Gili); elio.musacchio@phd.unipi.it (E. Musacchio); m.nissim@rug.nl (M. Nissim);
viviana.patti@unito.it (V. Patti); matteo.rinaldi@unito.it (M. Rinaldi); d.scalena@campus.unimib.it (D. Scalena)
🌐 <https://gattanasio.cc/> (G. Attanasio);
<https://swap.di.uniba.it/members/basile.pierpaolo/> (P. Basile);
<https://github.com/crux82> (D. Croce); <https://github.com/rosakun> (M. Francis); <https://github.com/Jj-source> (J. Gili);
<https://github.com/m-elio> (E. Musacchio);
<https://malvinanissim.github.io> (M. Nissim);
<https://github.com/vivpatti> (V. Patti); <https://github.com/mrinaldi97> (M. Rinaldi); <https://github.com/DanielSc4> (D. Scalena)
🆔 0000-0001-6945-3698 (G. Attanasio); 0000-0002-0545-1105 (P. Basile); 0009-0000-0193-2131 (F. Borazio); 0000-0001-9111-1950 (D. Croce); 0009-0007-7638-9963 (M. Francis); 0009-0007-1343-3760 (J. Gili); 0009-0006-9670-9998 (E. Musacchio); 0000-0001-5289-0971 (M. Nissim); 0000-0001-5991-370X (V. Patti); 0009-0004-7488-8855 (M. Rinaldi); 0009-0006-0518-6504 (D. Scalena)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Introduction

In parallel with the ongoing and constant development of new Large Language Models (LLMs), it has increased the need for understanding their abilities, how they differ from one another, and how they improve compared to previous iterations. To meet this need, the last couple of years have witnessed multiple efforts to put together new—or revisiting existing—benchmarks against which the performance and progress of LLMs can be monitored. These benchmarks include different tasks to test a variety of characteristics and abilities that are assumed to be associated with LLMs at different degrees. To mention a few, these span from multiple-choice questions of various sorts, commonsense and mathematical reasoning, and a variety of linguistic phenomena. BIG-bench [1] is currently the largest and most comprehensive benchmark, including over 200 tasks, almost all in English, which have been collaboratively contributed by researchers across the globe.

However, benchmarking progress for languages other than English has not improved with comparable quality. In many cases, evaluation datasets are automatic translations of their English counterparts, yielding not only a less native and possibly ungrammatical language but also

a cultural picture that is distant from the target language.

In the Italian NLP landscape, there is a long tradition of evaluation through the contribution of shared tasks. These benchmarks have been collected and run for almost 20 years in the context of the EVALITA campaigns (<https://www.evalita.it/>). The campaigns have fostered the creation of training and evaluation resources and models natively developed for Italian. Based on such resources, UINAUL (Unified Interactive Natural Understanding of the Italian Language)[2], an integrated benchmark for Italian NLU including six tasks has been recently proposed, and tested with available Italian and multilingual language models.

Except for CHANGE-IT [3], a generation task focused on headline transformation and organized within the EVALITA 2020 edition, all EVALITA tasks have focused on classification problems (some have been recast as generation problems as part of a resource release within the “Risorse per la Lingua Italiana” (RiTA) community [4]). However, to improve upon existing benchmarks, we wanted the core of a dynamic reference benchmark for Italian to include new tasks specifically focused on testing LLMs’ abilities.

Therefore, in the steps of this solid Italian benchmarking tradition, and in line with the most recent developments regarding the evaluation of LLMs, AILC—the Italian Association for Computational Linguistics—has launched “Challenge the Abilities of LLanguage Models in ITALian” (CALAMITA), a large-scale collaborative initiative across the whole Italian NLP community to develop a dynamic and growing benchmark for evaluating LLMs’ capabilities in Italian. This strategy would ensure a high diversity of tasks and, thus, of tested capabilities. It would distribute the effort of creative resources natively in Italian across many researchers and practitioners.

In the long term, we aim to establish a continuously growing suite of tasks that can be accessed through a shared platform and a live leaderboard so that any newly developed LLM, either multilingual or Italian monolingual, can be readily assessed. In the short term, we have started to build the CALAMITA benchmark through a series of challenges collaboratively contributed by the research community (Section 2). Also, we have established an evaluation framework that enables running the current and possibly future challenges in a centralized and coherent manner. This short paper summarises the collaborative procedure, the challenges currently included in CALAMITA¹, and the evaluation procedure.

2. Collaborative Methodology

The CALAMITA approach is inspired by standard Natural Language Processing shared tasks, giving the benchmark

¹The CALAMITA website: <https://clic2024.ilc.cnr.it/calamita/>.

a strong collaborative nature. The Italian Association for Computational Linguistics (AILC, <https://www.ai-lc.it>) launched a public call, mainly aimed at the Italian NLP community but spread across the standard international communication channels, asking for challenges and corresponding datasets, that LLMs could be tested on.

Participants contributing to a challenge were expected to provide an explanation and motivation for a given challenge, as well as a dataset that reflects that challenge. It was also asked to provide any information relevant to the dataset (provenance, annotation, distribution of labels or phenomena, etc.) Evaluation metrics and examples were also expected, along with the task and dataset submission. Existing relevant datasets could also be submitted as long as they made an interesting contribution to the benchmark and were natively created in Italian. To standardize the contribution to the CALAMITA benchmark, all proposed tasks with existing or new datasets had to follow a predefined template created and distributed by the CALAMITA organizers.

Creating the CALAMITA benchmark and the first round of LLM evaluation required several steps. In the first phase, all prospective participants submitted a pre-proposal. In case of a positive evaluation, based on compliance with the requirements and balance across submissions – participants were then asked to submit the final and complete challenge, following the provided CALAMITA template, in phase two. A final report was also requested for each accepted task, providing information on implementing the code for the evaluation.

The data and evaluation team set up the final CALAMITA benchmark by compiling the data and code of all the proposed tasks. We forked the Language Model Evaluation Harness tool² to create a custom CALAMITA version by including all the accepted tasks. Once the benchmark was assembled, the CALAMITA organizers ran zero- or few-shot experiments with a selection of LLMs. No tuning materials or experiments are expected at this project stage. Also, while we expect that CALAMITA, in the longer run, will be further populated by additional tasks and will have its own publicly accessible leaderboard, allowing for model testing, in this first stage, the choice of LLMs to be evaluated and the evaluation procedure is centralized.

3. Challenges

The preliminary call for tasks yielded the submission of over 20 proposals. Almost all of them were retained and are part of the present CALAMITA challenge, apart from the proposals that aimed at testing abilities that LLMs should not be expected to have, such as abilities typical of information retrieval engines and the proposals that

²<https://github.com/EleutherAI/lm-evaluation-harness>









Ability tested	Description	Count
 Commonsense knowledge	General knowledge about the world that is typically taken for granted in everyday life, e.g., everyday cause-and-effect relationships, situational judgments, physical properties, and basic social interactions.	19
 Factual knowledge	Knowledge of concrete, verifiable facts about the world, e.g., definitions, historical events, or scientific concepts.	12
 Linguistic knowledge	Linguistically motivated tasks that test specific language skills, e.g., word sense disambiguation, coreference resolution, or acceptability judgment.	22
 Formal reasoning	Ability to understand and use formally logical principles to solve problems, e.g., mathematical problems.	9
 Fairness and bias	Evaluates a model’s capacity to handle sensitive tasks, including exclusive and stereotyped language understanding and detecting offensive or biased language towards social groups.	6
 Code generation	Ability to generate fully functioning code for a specific programming language.	1
 Machine translation	Ability to translate a sentence from a source language into another language, with one of the two being Italian.	2
 Summarization	Ability to create relevant summaries of a given excerpt, e.g., news headline generation or news reduction.	2

Table 1

Categories of abilities tested by CALAMITA tasks. Tasks test general abilities such as knowledge about true facts, commonsense, and logical reasoning (top) or specific NLP-oriented abilities such as code generation or machine translation (bottom). Each task may require models to exhibit more than one ability.

required manual evaluation. In what follows, we briefly describe each task included in CALAMITA and refer the reader to each of the challenges’ reports for further details. In Table 1, we describe the macro categories under which the CALAMITA tasks can be grouped, where categories are broad classes of tested abilities. Table 2 shows which abilities apply to each challenge.

ABRICOT (ABstRactness and Inclusiveness in CONtext) [5] is a task designed to evaluate Italian language models on their ability to understand and assess the abstractness and inclusiveness of language, two nuanced features that humans naturally convey in everyday communication. Unlike binary categorizations such as abstract/concrete or inclusive/exclusive, these features exist on a continuous spectrum with varying degrees of intensity. The task is based on a manual collection of sentences that present the same noun phrase (NP) in different contexts, allowing its interpretation to vary between the extremes of abstractness and inclusiveness. This challenge aims to verify how LLMs perceive subtle linguistic variations and their implications in natural language.

AMELIA (Argument Mining Evaluation on Legal documents in ItAlian) [6] is a challenge consisting of three classification tasks in the context of argument mining in the legal domain. The tasks are based on a dataset of 225 Italian decisions on Value Added Tax, annotated to identify and categorize argumentative text. The objective of the first task is to classify each argumen-

tative component as a premise or conclusion. In contrast, the second and third tasks aim at classifying the type of premise: legal vs factual, and its corresponding argumentation scheme. The classes are highly unbalanced, hence evaluation is based on the macro F1 score.

BEEP (BEst DrivEr’s License Performer) [7] is a benchmark to evaluate large language models in the context of a simulated Italian driver’s license exam. This challenge tests the models’ ability to understand and apply traffic laws, road safety regulations, and vehicle-related knowledge through a series of true/false questions. The dataset is derived from official ministerial materials used in the Italian licensing process, explicitly targeting Category B licenses.

BLM-It (Blackbird Language Matrices) [8] is a task made of linguistic puzzles (matrices) around language-related problems, focusing on formal and semantic properties of language. A BLM matrix consists of a context set and an answer set. The context is a sequence of sentences that encodes implicitly an underlying generative linguistic rule. The contrastive multiple-choice answer set includes negative examples following corrupted generating rules. The models are prompted in a few-shot setting. The datasets comprise a few prompts for a few-shot setting.

DIMMI (Drug InforMation Mining in Italian) [9] is a task aimed at evaluating the proficiency of Large

Language Models in extracting drug-specific information from Patient Information Leaflets. The challenge evaluates the effectiveness of processing complex medical information in Italian and is approached as an information extraction task in a zero-shot setting, based on the model’s pre-existing knowledge or through in-context learning. Evaluation is performed against a manually created gold standard.

ECWCA (Educational CrossWord Clues Answering) [10] is designed to evaluate the knowledge and reasoning capabilities of LLMs through crossword clue-answering. The challenge consists of two tasks: a standard question-answering format where the LLM is asked to solve crossword clues and a variation where the model is given hints about the word lengths of the answers, which is expected to help models with reasoning abilities.

EurekaRebus [11] is a task that tests the ability of LLMs to conduct multi-step, knowledge-intensive inferences while respecting predefined constraints. LLMs are prompted to reason step-by-step to solve verbalized variants of rebus games. Verbalized rebuses replace visual cues with crossword definitions to create an encrypted first pass, making the problem entirely text-based. Multiple metrics are used to grasp the models’ performance in knowledge recall, constraints adherence, and re-segmentation abilities across reasoning steps.

GATTINA (GenerAtion of TiTles for Italian News Articles) [12] is a task that aims to assess the ability of LLMs to generate headlines for science news articles. Aspects such as the appropriateness of the summary, creativity, and attractiveness are evaluated through a battery of metrics. The benchmark consists of a large dataset of science news articles and their corresponding published headlines from ANSA Scienza and Galileo, two prominent Italian media outlets.

GEESE (Generating and Evaluating Explanations for Semantic Entailment) [13] is focused on evaluating the impact of generated explanations on the predictive performance of language models for the task of Recognizing Textual Entailment in Italian. Using a dataset enriched with human-written explanations, two large language models are employed to generate and utilize explanations for semantic relationships between sentence pairs. GEESE assesses the quality of generated explanations by measuring changes in prediction accuracy when explanations are provided.

GFG (Gender-Fair Generation) [14] is a task designed to assess and monitor the recognition and generation of gender-fair language in both mono- and cross-

lingual scenarios. It includes three tasks: (1) the detection of gender-marked expressions in Italian sentences, (2) the rewriting of gendered expressions into gender-fair alternatives, and (3) the generation of gender-fair language in automatic translation from English to Italian. The challenge relies on three different annotated datasets: the GFL-it corpus, which contains Italian texts extracted from administrative documents provided by the University of Brescia; GeNTE, a bilingual test set for gender-neutral rewriting and translation built upon a subset of the Europarl dataset; Neo-GATE, a bilingual test set designed to assess the use of non-binary neomorphemes in Italian for both fair formulation and translation tasks.

GITA (Graded Italian Annotated Dataset) [15] investigates the physical commonsense reasoning capabilities of large language models, assessing their low-level understanding of the physical world using a test set in the Italian language. Three specific tasks are evaluated: identifying plausible and implausible stories within our dataset, identifying the conflict that generates an implausible story, and identifying the physical states that make a story implausible. It is written and annotated by a professional linguist.

INVALSI [16] is a benchmark based on the Invalsi tests administered to students within the Italian school system. Expert pedagogists prepare these tests with the explicit goal of testing average students’ performance over time across Italy. There are two benchmarks: Invalsi MATE (420 questions), which targets the models’ performance on mathematical understanding, and Invalsi ITA (1279 questions), which evaluates language understanding in Italian.

ITA-SENSE (ITALian word SENSE disambiguation) [17] is a task that assesses LLMs’ abilities in understanding lexical semantics through Word Sense Disambiguation. The classical Word Sense Disambiguation task is cast as a generative problem formalized as two tasks: [T1] Given a target word and a sentence in which the word occurs, generate the correct meaning definition; [T2] Given a target word and a sentence in which the word occurs, choose the correct meaning definition from a predefined set. For CALAMITA, LLMs are tested in a zero-shot setting.

MACID (Multimodal ACTION IDentification) [18] is a task aimed at evaluating LLMs to differentiate between closely related action concepts based on textual descriptions alone. The challenge is inspired by the “find the intruder” task, where models must identify an outlier among a set of 4 sentences that describe similar yet distinct actions. The dataset highlights action-predicate

mismatches, where the same verb may describe different actions, or different verbs may refer to the same action. Although mono-modal (text-only), the task is designed for future multimodal integration, linking visual and textual representations to enhance action recognition.

MT (Machine Translation) [19] is a task that aims at testing the ability of LLMs in automatic translation, focusing on Italian and English (in both directions). The task proposes a benchmark composed of two datasets covering different domains and with varying distribution policies. Performances are reported in terms of four evaluation metrics, whose scores allow an overall evaluation of the quality of the automatically generated translations.

Multi-IT [20] is a large-scale Multi-Choice Question Answering (MCQA) dataset for evaluating the factual knowledge and reasoning abilities of LLMs in Italian. This contribution aims to counteract the disadvantages of using MCQA benchmarks that are automatically translated from English and may sound unnatural, contain errors, or use linguistics constructions that do not align with the target language. In addition, they may introduce topical and ideological biases reflecting Anglo-centric perspectives. Multi-IT comprises over 110,000 manually written questions sourced directly from preparation quizzes for Italian university entrance exams or for exams for public sector employment in Italy.

Pejorativity [21] is a task to investigate misogyny expressed through neutral words that can assume a negative connotation when functioning as pejorative epithets. This challenge addresses a) the disambiguation of such ambiguous words in a given context; b) the detection of misogyny in instances that contain such polysemic words. The task is divided into two parts, both framed as a binary classification. In Task A, the model is asked to define if, given a tweet, the target word is used in a pejorative or non-pejorative way. In Task B, the model is asked whether the whole sentence is misogynous.

PERSEID (PERSpEctivist Irony Detection) [22] considers the task of irony detection from short social media conversations collected from Twitter (X) and Reddit. Data is leveraged from MultiPICO, a recent multilingual dataset with disaggregated annotations and annotators' metadata. The dataset evaluates whether prompting LLMs with additional annotators' demographic information (gender only, age only, and the combination of the two) improves performance compared to a baseline in which only the input text is provided.

TRACE-it (Testing Relative clAuses Comprehension through Entailment in Italian) [23] is a benchmark

designed to evaluate the ability of LLMs to comprehend a specific type of complex syntactic construction in Italian: object relative clauses. The challenge is framed as a binary entailment task where, given a complex sentence, the model is tasked with determining whether it logically entails a simpler yes/no implication.

Termite [24] focuses on the Text-to-SQL task in Italian. Natural language queries are written natively in Italian, and the models are expected to turn them into SQL queries. The dataset is built to be invisible to search engines since it is locked under an encryption key delivered along the resource to reduce accidental inclusion in upcoming training sets. It contains hand-crafted databases in different domains, each with a balanced set of NL-SQL query pairs. The NL questions are built in such a way that they can be solved by a model relying only on its linguistic proficiency and an analysis of the schema, with no external knowledge needed.

VeryfIT [25] is designed to evaluate the in-memory factual knowledge of language models on data written by professional fact-checkers, posing it as a true or false question. Topics of the statements vary, but most are in specific domains related to the Italian government, policies, and social issues. The task presents several challenges: extracting statements from segments of speeches, determining appropriate contextual relevance both temporally and factually, and verifying the statements' accuracy.

ItaEval [26] is a multifaceted evaluation suite comprising three overarching task categories: (i) natural language understanding, (ii) commonsense and factual knowledge, and (iii) bias, fairness, and safety [4]. ItaEval is a collection of 18 tasks encompassing existing and new datasets. The so-compiled ItaEval suite provides a standardized, multifaceted framework for evaluating Italian language models, facilitating more rigorous and comparative assessments of model performance.

4. Evaluation Strategy

Rooted in its very nature, CALAMITA's biggest challenge is standardizing evaluation across many tasks and scenarios. To account for such high variability, we settled on a few fundamental choices that shape CALAMITA's core principles (**Design choices**) and left broad freedom to challenge participants to specify fine-grained aspects of their tasks (**Participant choices**). Base design choices shared across all tasks and high task-specific customization balance standardization and versatility.





Task									Type
ABRICOT			✓						
AMELIA	✓	✓							
BEEP		✓							
BLM-It			✓						
DIMMI		✓*							
ECWCA	✓	✓	✓	✓					
EurekaRebus	✓	✓	✓	✓					
GATTINA									✓
GEESE	✓			✓					
GFG			✓		✓		✓		
GITA	✓			✓					
INVALSI	✓	✓	✓	✓					
ITA-SENSE			✓						
MACID	✓		✓						
MT									✓
Mult-IT	✓	✓	✓	✓					
PejorativITy			✓		✓				
PERSEID	✓		✓						
Termite								✓	
TRACE-it			✓	✓					
VeryfIT		✓	✓						
ItaEval									
ItaCoLA			✓						
Belebele-it		✓*							
News-Sum									✓
IronITA	✓		✓						
SENTIPOLC	✓		✓						
SQuAD-it		✓*							
TruthfulQA-it		✓							
ARC-it	✓	✓	✓	✓					
XCOPA-it	✓		✓		✓				
HellaSwag-it	✓				✓				
AMI	✓		✓			✓			
HONEST	✓**								
GeNTE rephrasing	✓		✓			✓			
Multilingual HateCheck	✓**		✓			✓			
HaSpeeDe2	✓		✓			✓			

Table 2

Abilities tested by each task in CALAMITA. *: task that require *contextualized* factual knowledge, e.g., reading comprehension tasks. **: tasks that require *stereotypical* commonsense knowledge, e.g., understanding the concept of misogyny.

Design choices. Following recent practices for language model evaluation [e.g., 27, 28], we consider every received task as a downstream task to be solved via standard prompting. We support two types of tasks: Multiple-Choice (MC) and Open-Ended (OE) generation. MC tasks require a model to pick one or more correct answers from a finite set. OE tasks require models to generate output tokens until a stopping criterion is met. For evaluating multiple-choice tasks, we rank all candidates by their likelihood conditioned on the prompt and pick the highest [29]. We normalize each option probability by the number of tokens. Closed-question question-answering is an example of an MC task. We do not adopt a single strategy for OE tasks, as evaluation depends on the semantics of the output. Machine translation and summarization are examples of OE tasks. Moreover, we standardize the decoding strategy across OE tasks. We use beam search ($n = 5$) for machine translation and greedy decoding for all other tasks. See Appendix A for the complete details.

To foster reproducibility, we base CALAMITA’s codebase on open-source tools. We forked and built our evaluation code upon *lm-eval* [30]. When possible, we recommended public and accessible data release to the participants through the HuggingFace Hub.³ We release our evaluation code at <https://github.com/CALAMITA-AILC/lm-evaluation-harness>.

Participant choices. In addition to the data associated with the task and the type (MC or OE), we request that each participating team provides specifics regarding compiling an arbitrary prompt and evaluating an arbitrary model generation. Among prompting details, task proposers specified a prompt template and the number of task demonstrations (0 for zero-shot, N for N-shot prompting). In few-shot cases, we requested where to sample the demonstrations and the sampling strategy (static, dynamic-random, or dynamic-sequential). Among the evaluation details, we requested that participants specify any post-processing function for model raw outputs, one or more evaluation metrics, and relative information. For reporting purposes, we collected a single evaluation score (the first metric listed by proposers).

Crucially, we relied upon meta-description and code to streamline the communication between the task proposers and the challenge organizers. Participants were tasked to provide such information through a single file following a set of guidelines.⁴

Model Selection. We tested Llama 3.1 8B Instruct [31] and ANITA [32], two state-of-the-art decoder-only lan-

guage models. Llama’s 3.1 variant introduces multilingual support to the family’s previous iteration. ANITA is a fine-tuned version of Llama 3 specializing in English and Italian tasks.

Our choice was driven by three primary reasons. First, both models are open-weight, well-known within the Italian NLP community, and explicitly support the Italian language. Second, they have been instruction fine-tuned, a training step that facilitates addressing tasks in zero-shot. Third, they are within the 8 billion parameter range, which allows for fast iteration and good performance.

Results. At the time of writing, some of the results are still being collected. To provide a comprehensive and dynamic overview, we refer the reader to the external page where they get regularly updated: <https://calamita-aile.github.io/calamita2024/>.

5. Limitations

CALAMITA is not intended to be an exhaustive benchmark for testing abilities of Italian LLMs, especially at this first release. Considering the strong collaborative nature of this benchmark, coherence across tasks might not be optimal, in spite of the efforts put in by the organisers to uniform all datasets and the evaluation procedure. Although we have paid attention to this issue, we cannot be absolutely certain that none of the datasets, in one form or another, have ended up in some training set, already.

Acknowledgments

The ItaEval tasks submitted to CALAMITA are the result of a joint effort of members of the “Risorse per la Lingua Italiana” community (rita-nlp.org): we thank every member who dedicated their time to the project. For providing the computational resources we thank CINECA (ISCRA grant: HP10C3RW9F; ISCRA C grant: CALAMITA – HP10CKZDYT), the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster and University of Turin for providing access to the HPC4AI cluster [33]. Malvina Nissim’s work is also part of the “Humane AI” theme of the Dutch Sectorplan for the Humanities. The work of Viviana Patti was partially supported by “HARMONIA” project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme. The work by Giuseppe Attanasio was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020. The work by Pierpaolo Basile and Elio

³Resulting from the effort for CALAMITA, 35 new datasets have been released with a permissive license.

⁴See the guidelines at <https://github.com/CALAMITA-AILC/calamita2024> and the information file at <https://gist.github.com/g8a9/f5e82d38ce12831323b20dc79b0452c9>

Musacchio was supported by the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU. The work of Matteo Rinaldi and Jacopo Gili has been partly supported by the Spoke “Future HPC & Big Data” of the ICSC - Centro Nazionale di Ricerca in “High Performance Computing, Big Data and Quantum Computing”, funded by European Union - NextGenerationEU.

References

- [1] A. Srivastava, D. Kleyjo, Z. Wu, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, *Transactions on Machine Learning Research* (2023).
- [2] V. Basile, L. Bioglio, A. Bosca, C. Bosco, V. Patti, UINAUIL: A unified benchmark for Italian natural language understanding, in: D. Bollegala, R. Huang, A. Ritter (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 348–356. URL: <https://aclanthology.org/2023.acl-demo.33>. doi:10.18653/v1/2023.acl-demo.33.
- [3] L. De Mattei, M. Cafagna, A. AI, F. Dell’Orletta, M. Nissim, A. Gatt, Change-it@ evalita 2020: Change headlines, adapt news, generate, EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020 (2020) 235.
- [4] G. Attanasio, P. Delobelle, M. La Quatra, A. Santilli, B. Savoldi, Itaeval and tweetyita: A new extensive benchmark and efficiency-first language model for italian, in: *CLiC-it 2024: Tenth Italian Conference on Computational Linguistics*, Date: 2024/12/04-2024/12/06, Location: Pisa, Italy, 2024.
- [5] G. Puccetti, C. Collacciani, A. A. Ravelli, A. Esuli, M. Bolognesi, ABRICOT - ABstRactness and Inclusiveness in CoNtExT: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [6] G. Grundler, A. Galassi, P. Santin, A. Fidelangeli, F. Galli, E. Palmieri, F. Lagioia, G. Sartor, P. Torroni, AMELIA - Argument Mining Evaluation on Legal documents in ItAlian: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [7] F. Mercorio, D. Poterti, A. Serino, A. Seveso, BEEP - BEst DrivEr’s License Performer: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [8] C. Jiang, G. Samo, V. Nastase, P. Merlo, BLM-It - Blackbird Language Matrices for Italian: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [9] R. Manna, M. P. Di Buono, L. Giordano, DIMMI - Drug InforMation Mining in Italian: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [10] A. Zugarini, K. Zeinalipour, A. Fusco, A. Zanollo, ECWCA - Educational CrossWord Clues Answering A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [11] G. Sarti, T. Caselli, A. Bisazza, M. Nissim, EurekaRebus - Verbalized Rebus Solving with LLMs: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [12] M. Francis, M. Rinaldi, J. Gili, L. De Cosmo, S. Iannaccone, M. Nissim, V. Patti, GATTINA - Generation of TiTles for Italian News Articles: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [13] A. Zaninello, B. Magnini, GEESE - Generating and Evaluating Explanations for Semantic Entailment: a CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [14] S. Frenda, A. Piergentili, B. Savoldi, M. Madeddu, M. Rosola, S. Casola, C. Ferrando, V. Patti, M. Negri, L. Bentivogli, GFG - Gender-Fair Generation: A CALAMITA Challenge, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [15] G. Pensa, E. Azurmendi, J. Etxaniz, B. Altuna,

- I. Gonzalez-Dios, GITA4CALAMITA - Evaluating the Physical Commonsense Understanding of Italian LLMs in a Multi-layered Approach: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [16] G. Puccetti, M. Cassese, A. Esuli, INVALSI - Mathematical and Language Understanding in Italian: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [17] P. Basile, E. Musacchio, L. Siciliani, ITA-SENSE - Evaluate LLMs' ability for ITALian word SENSE disambiguation: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [18] A. A. Ravelli, R. Varvara, L. Gregori, MACID - Multimodal ACTION IDentification: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [19] M. Cettolo, A. Piergentili, S. Papi, M. Gaido, M. Negri, L. Bentivogli, MAGNET - MACHines GENerating Translations: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [20] M. Rinaldi, J. Gili, M. Francis, M. Goffetti, V. Patti, M. Nissim, Mult-IT Multiple Choice Questions on Multiple Topics in Italian: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [21] A. Muti, PejorativITy - In-Context Pejorative Language Disambiguation: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [22] V. Basile, S. Casola, S. Frenda, S. M. Lo, PERSEID - Perspectivist Irony Detection: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [23] D. Brunato, TRACE-it: Testing Relative cLAuses Comprehension through Entailment in ITALian: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [24] F. Ranaldi, E. S. Ruzzetti, D. Onorati, F. M. Zanzotto, L. Ranaldi, Termite Italian Text-to-SQL: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [25] J. Gili, V. Patti, L. Passaro, T. Caselli, VeryFIT - Benchmark of Fact-Checked Claims for Italian: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [26] G. Attanasio, M. La Quatra, A. Santilli, B. Savoldi, ItaEval: A CALAMITA Challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [27] S. Mehta, M. H. Sekhavat, Q. Cao, M. Horton, Y. Jin, C. Sun, S. I. Mirzadeh, M. Najibi, D. Belenko, P. Zatloukal, et al., OpenELM: An efficient language model family with open training and inference framework, in: Workshop on Efficient Systems for Foundation Models II@ ICML2024, 2024.
- [28] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang, et al., OLMo: Accelerating the science of language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15789–15809. URL: <https://aclanthology.org/2024.acl-long.841>. doi:10.18653/v1/2024.acl-long.841.
- [29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Teusz Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NeurIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/

- file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [30] S. Biderman, H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, et al., Lessons from the trenches on reproducible evaluation of language models, arXiv preprint arXiv:2405.14782 (2024).
- [31] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783 (2024).
- [32] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, arXiv preprint arXiv:2405.07101 (2024).
- [33] M. Aldinucci, S. Rabellino, M. Pironti, F. Spiga, P. Viviani, M. Drocco, M. Guerzoni, G. Boella, M. Mellia, P. Margara, I. Drago, R. Marturano, G. Marchetto, E. Piccolo, S. Bagnasco, S. Lusso, S. Vallero, G. Attardi, A. Barchiesi, A. Colla, F. Galeazzi, Hpc4ai, an ai-on-demand federated platform endeavour, in: ACM Computing Frontiers, Ischia, Italy, 2018. URL: https://iris.unito.it/retrieve/handle/2318/1765596/689772/2018_hpc4ai_ACM_CF.pdf. doi:10.1145/3203217.3205340.

A. Experimental Details

A.1. Technical Details

We run our experiments on the LEONARDO HPC infrastructure (Booster partition). The booster module partition is based on BullSequana XH2135 supercomputer nodes, each with four NVIDIA Tensor Core GPUs (custom Ampere A100 GPU 64GB HBM2e, NVLink 3.0 (200GB/s)) and a single Intel CPU.⁵

We forked the `lm-eval-harness` official repository at the commit with hash `b2bf7bc4a601c643343757c92c1a51eb69caf1d7`. We report all technical details on our official webpage.⁶

A.2. Generation Configuration

Table 3 reports the generation parameters we used for Open-Ended tasks.

Parameter	Value
Batch size	1*
Temperature	0.0
Sampling	False
Stopping criteria	<code>\n\n, </s>, < im_end >, ". ", < eot_id >, < end_of_text ></code>

Table 3

Generation Parameters. *: we set beam search to 5 for machine translation tasks.

⁵<https://www.hpc.cineca.it/systems/hardware/leonardo/>

⁶<https://calamita-aile.github.io/calamita2024/>