Challenging the Abilities of Large Language Models in Italian: a Community Initiative

Malvina Nissim*[22], Danilo Croce**[27], Viviana Patti†[31], Pierpaolo Basile‡[18], Giuseppe Attanasio[11], Elio Musacchio[18,26], Matteo Rinaldi[31], Federico Borazio[27], Maria Francis[22,30], Jacopo Gili[31], Daniel Scalena[22,23], Begoña Altuna[29], Ekhi Azurmendi[29], Valerio Basile[31], Luisa Bentivogli[6], Arianna Bisazza[22], Marianna Bolognesi[19], Dominique Brunato[10], Tommaso Caselli[22], Silvia Casola[15], Maria Cassese[12], Mauro Cettolo[6], Claudia Collacciani[19], Leonardo De Cosmo[2], Maria Pia Di Buono[24], Andrea Esuli[12], Julen Etxaniz[29], Chiara Ferrando[31], Alessia Fidelangeli[19], Simona Frenda[8], Achille Fusco[13], Marco Gaido[6], Andrea Galassi[19], Federico Galli[19], Luca Giordano[24], Mattia Goffetti[1], Itziar Gonzalez-Dios[29], Lorenzo Gregori[20], Giulia Grundler[19], Sandro Iannaccone[7], Chunyang Jiang[9,21], Moreno La Quatra[14], Francesca Lagioia[4,19], Soda Marem Lo[31], Marco Madeddu[31], Bernardo Magnini[6], Raffaele Manna[24], Fabio Mercorio[23], Paola Merlo[9,21], Arianna Muti[3], Vivi Nastase[9], Matteo Negri[6], Dario Onorati[27], Elena Palmieri[19], Sara Papi[6], Lucia Passaro[26], Giulia Pensa[29], Andrea Piergentili[6,30], Daniele Potertì[23], Giovanni Puccetti[12], Federico Ranaldi[27], Leonardo Ranaldi[27], Andrea Amelio Ravelli[19,30], Martina Rosola[17], Elena Sofia Ruzzetti[27], Giuseppe Samo[9], Andrea Santilli[16], Piera Santin[19], Gabriele Sarti[22], Giovanni Sartor[4,19], Beatrice Savoldi[6], Antonio Serino[23], Andrea Seveso[23], Lucia Siciliani[18], Paolo Torroni[19], Rossella Varvara[25], Andrea Zaninello[6], Asya Zanollo[13], Fabio Massimo Zanzotto[27], Kamyar Zeinalipour[28], Andrea Zugarini[5]

[1] Alpha Test S.r.l.; [2] ANSA; [3] Bocconi University; [4] European University Institute; [5] Expert.ai; [6] Fondazione Bruno Kessler; [7] Galileo Net; [8] Heriot-Watt University; [9] Idiap Research Institute; [10] ILC-CNR; [11] Instituto de Telecomunicações; [12] ISTI-CNR; [13] IUSS Pavia; [14] Kore University of Enna; [15] LMU Munich; [16] Sapienza University of Rome; [17] Universitat de Barcelona; [18] University of Bari Aldo Moro; [19] University of Bologna; [20] University of Florence; [21] University of Geneva; [22] University of Groningen; [23] University of Milano-Bicocca; [24] University of Naples "L'Orientale"; [25] University of Pavia; [26] University of Pisa; [27] University of Rome Tor Vergata; [28] University of Siena; [29] University of the Basque Country (UPV/EHU); [30] University of Trento; [31] University of Turin.

The rapid progress of Large Language Models (LLMs) has transformed natural language processing and broadened its impact across research and society. Yet, systematic evaluation of these models, especially for languages beyond English, remains limited. "Challenging the Abilities of LAnguage Models in ITAlian" (CALAMITA) is a large-scale collaborative benchmarking initiative for Italian, coordinated under the Italian Association for Computational Linguistics. Unlike existing efforts that focus on leaderboards, CALAMITA foregrounds methodology: it federates

* Corresponding author: m.nissim@rug.nl
** Corresponding author: croce@info.uniroma2.it
† Corresponding author: viviana.patti@unito.it
‡ Corresponding author: pierpaolo.basile@uniba.it

more than 80 contributors from academia, industry, and the public sector to design, document, and evaluate a diverse collection of tasks, covering linguistic competence, commonsense reasoning, factual consistency, fairness, summarization, translation, and code generation. Through this process, we not only assembled a benchmark of over 20 tasks and almost 100 subtasks, but also established a centralized evaluation pipeline that supports heterogeneous datasets and metrics. We report results for four open-weight LLMs, highlighting systematic strengths and weaknesses across abilities, as well as challenges in task-specific evaluation. Beyond quantitative results, CALAMITA exposes methodological lessons: the necessity of fine-grained, task-representative metrics, the importance of harmonized pipelines, and the benefits and limitations of broad community engagement. CALAMITA is conceived as a rolling benchmark, enabling continuous integration of new tasks and models. This makes it both a resource – the most comprehensive and diverse benchmark for Italian to date – and a framework for sustainable, community-driven evaluation. We argue that this combination offers a blueprint for other languages and communities seeking inclusive and rigorous LLM evaluation practices.

1. Introduction and Motivation

Large Language Models (LLMs) have rapidly become a cornerstone of contemporary Natural Language Processing (NLP), demonstrating impressive abilities across a wide range of linguistic and more general downstream tasks. Concurrently, beyond their dominance in NLP, LLMs have now become pervasive in the daily life of many, and also in the research activities of scholars of a large variety of fields.

Evaluating where these models excel and, crucially, where they fall short is therefore not only key to a better understanding of current developments in NLP and thus better guidance for future steps, but also a fundamental contribution to a more aware, responsible, and effective use of such models in research and society at large.

How to do model benchmarking in practice, however, remains a significant challenge. The question of which benchmarks to use to assess the acquired abilities of LLMs is far from trivial, especially when considering languages beyond English, Chinese, and programming code, which together dominate the training data of most modern LLMs (for instance, less than 0.1% of training examples in models such as LLaMA2 are in Italian (Touvron et al. 2023)). The crux of the issue is not the inherent complexity of languages like Italian, but rather the limited international attention devoted to their evaluation, and the effort required to run language-specific campaigns. At large scale, the common trend to date for most languages other than English has been to rely on translated benchmarks (Thellmann et al. 2024; Xuan et al. 2025) or synthetic data approaches that fail to capture the true linguistic challenges faced by models in real-world scenarios and the true abilities of models in those languages (Plaza et al. 2024; Wu et al. 2025). Without native, authentic, human-crafted evaluation datasets, both fine-tuning and robust assessment of LLMs in underrepresented languages become academic exercises aimed at extracting leaderboard scores with no real understanding of the actual abilities of models in those languages and the cultures they express.

Faced with this state of affairs, and drawing on a long-standing tradition of Italian-focused evaluation campaigns (Attardi et al. 2015; Basile et al. 2017; Passaro et al. 2020; Basile et al. 2022), the Italian NLP community has mobilized. Under the auspices of AILC (Associazione Italiana di Linguistica Computazionale)[1], the CALAMITA (*Chal-*

1 https://www.ai-lc.it/en/

lenge the Abilities of Language Models in Italian) initiative was launched as a collective, grassroots effort to create high-quality and broad-interest evaluation data for Italian LLMs and set up a dynamic benchmarking framework for the benefit of the research community at large. Leveraging AILC's expertise in NLP and evaluation campaigns, CALAMITA brings together the necessary sensitivity for language-sensitive challenges, the capacity to automate and centralize large-scale evaluation, and the access to national high-performance computing resources such as the Leonardo supercomputer[2], which remains out of reach for many individual researchers.

The community response has been extraordinary. More than 80 contributors from 31 different institutions both within and beyond NLP, including non-academic stakeholders such as journalists, publishers, and public sector professionals participated in defining tasks, annotating data, and supporting the automation of evaluation pipelines and the use of computational clusters. This distributed, self-organized effort enabled the evaluation of four LLMs across 22 distinct challenges and 95 sub-tasks, all orchestrated in a collaborative and centralized fashion. It also established an infrastructure which allows for the continuous addition of more challenges to the dynamic benchmark and the evaluation of other models (see Section 5.2).

Generating a large volume of evaluation results presents its own challenges. Within CALAMITA we have organized the challenges in a taxonomy to better see generalizations of model behaviors, while we have also encouraged the development of task-representative metrics so that model performance can be studied at a more fine-grained level without reducing it to a single averaged score. Indeed, CALAMITA does not aim to simply provide an abstract ranking of LLMs, nor to declare a definitive "winner" among models. Instead, it reflects the unique experience of a research association that has succeeded in federating multiple and diverse groups towards the creation of a unique resource. This resource is not limited to the current benchmark, which is composed of a large diversity of tasks, and has enabled a first round of experiments enriched with error analysis — this resource is the methodology itself: a collaborative framework for dataset creation and evaluation that combines linguistic insight, computational efficiency, and community engagement.

CALAMITA is not conceived as a one-off experiment, but as an ongoing, sustainable initiative, grounded on a rolling mechanism which will enable the continuous addition of tasks and evaluation of models. Its dynamics will continue to be community-driven, and guided by the constantly evolving needs of the Italian, as well as the international research landscape. Since this framework has proven successful in promoting a robust and meaningful evaluation of LLMs in languages beyond the current mainstream, we hope that our experience can serve as a blueprint for other research communities towards a more inclusive and comprehensive landscape for natural language processing research. As multilingual LLMs continue to advance, it is crucial to ensure that evaluation practices keep pace, especially for languages with limited resources and less international visibility.

For the reasons above, this paper contains not only a description of the challenges composing the CALAMITA benchmark at the time of writing, and results – both quantitative and qualitative – on four open-weight LLMs, but also a detailed explanation of the collaborative methodology we adopted both for data creation and for model evaluation. We discuss the linguistic, practical, and technical challenges encountered, and the solutions developed collaboratively by the community. Finally, we outline

2 https://leonardo-supercomputer.cineca.eu/

ongoing and future directions, inviting continued participation and replication of our approach in other contexts.

2. Related Work

Benchmarking has long served as a key contributor to empirical progress in NLP. In the past, many benchmarks were designed to test the general linguistic competence of language models through tasks such as summarization or paraphrase detection, which served as proxies for broader language understanding. Others focused on more narrowly defined tasks with a view to downstream applications, such as review classification, or hate speech detection, with datasets often designed both for model development and evaluation. This was reflected in long-standing evaluation campaigns such as SemEval[3], or–in the Italian context–EVALITA[4].

However, the rapid improvement of large language models (LLMs) is currently shifting the benchmarking landscape. More than focusing on downstream applications with models trained to solve specific tasks, there is a growing need for benchmarks that help to capture the diversity of LLMs' emerging abilities, especially in light of the fact that they have not been explicitly trained for those. Example abilities are reasoning, factual consistency, behavioural characteristics, and alignment with human values. This section does not attempt to be a comprehensive survey of the literature on language model evaluation. Instead, in the first part we illustrate the diversity of LLM benchmarks and the types of tested abilities they include; in the second we discuss initiatives that are similar to ours in scope and methodological approach.

Diversity of tasks and tested abilities in recent benchmarking. Many benchmarks continue to focus on traditional language processing tasks, including question answering (Kwiatkowski et al. 2019; Joshi et al. 2017; Choi et al. 2018; Lin, Hilton, and Evans 2022), reading comprehension (Dua et al. 2019; Lai et al. 2017; Bandarkar et al. 2024), word prediction (Paperno et al. 2016), sentence completion (Zellers et al. 2019), summarization (Ladhak et al. 2020), paraphrase detection (Yang et al. 2019), acceptability judgment (Warstadt et al. 2020; Suijkerbuijk et al. 2025; Zhang et al. 2024e), and dialogue modeling (Bai et al. 2024a). Such foundational tasks are often grouped into large-scale benchmarks, such as GLUE (Wang et al. 2018), SuperGLUE (Wang et al. 2019) or XTREME (Hu et al. 2020), which have for years served as reference benchmarks for first generation pretrained models such as BERT, RoBERTa, and GPT-2. Differently from more recent ability-testing benchmarks, a feature of GLUE and XTREME is that they provide task-specific training datasets on which pretrained models should be fine-tuned, in addition to evaluation sets for testing.

Research into the emerging abilities of LLMs has attracted diverse lines of inquiry. We outline several below, including domain-specific knowledge, multiple types of reasoning, commonsense knowledge, cognitive aspects, and pragmatics, and fairness. These are also categories that we draw upon to characterize and classify our CALAMITA challenges (see Table 1 and Section 4.1).

Domain-specific benchmarks have been developed to test how well language models can operate within particular fields, either by assessing their factual knowledge or their ability to perform tasks that are part of real-world applications. Benchmarks have

3 https://semeval.github.io/
4 https://www.evalita.it/

been designed to assess knowledge in science (Welbl, Liu, and Gardner 2017; Guo et al. 2023) or finance (Krumdick et al. 2024). AGIEval (Zhong et al. 2024) compiles questions from standardized exams such as the SAT or legal entry exams. In professional contexts, ACI-Bench (Yim et al. 2023) evaluates models on clinical note generation for medical visits, and QMSum (Zhong et al. 2021) assesses automatic meeting summarization in multiple domains.

Recently, a central open question is the extent to which LLMs can perform formal reasoning – that is, the ability to reason abstractly according to explicit rules. Popular tasks targeting formal reasoning include the solution of mathematical word problems (Wang, Liu, and Shi (2017), Cobbe et al. (2021), Li et al. (2024), see Chen et al. (2022) for geometry problems) and code generation (Austin et al. 2021; Yan et al. 2024; Zhang et al. 2024a; Lu et al. 2021; Khan et al. 2024; Kasner and Dusek 2024), but benchmarks have also been introduced testing performance on algorithmic problems (Fan et al. 2024; Parmar et al. 2024), graph extraction (Du et al. 2024), quantitative reasoning (Ravichander et al. 2019) and fallacy detection (Helwe et al. 2024). Commonsense reasoning benchmarks have emerged as a way to evaluate forms of reasoning that are not governed by formal logic but are typically intuitive for humans (Clark et al. 2018; Ponti et al. 2020; Sakaguchi et al. 2021; Frohberg and Binder 2022). These tasks aim to assess the models' ability to make plausible inferences about everyday situations. Interestingly, what is regarded as commonsense may be culture-specific, and this understanding may be evaluated independently (Sun et al. 2024). Related areas such as physical commonsense (Bisk et al. 2020), temporal reasoning, (Zhang et al. 2024d), and natural language inference (Williams, Nangia, and Bowman 2018; Sadat and Caragea 2022) are often included under this umbrella. To assess various forms of reasoning at once, using language models to solve puzzles is a creative branch of research; for example, Todd et al. (2024) test lateral thinking by having LLMs solve The New York Times' *Connections* puzzles. A growing area of evaluation builds on insights from cognitive science and psychology to probe the extent to which LLMs exhibit human-like social and pragmatic reasoning. Theory-of-Mind (ToM) benchmarks test whether models can infer the mental states of others – a fundamental component of social intelligence (Kosinski 2023; Hu, Sosa, and Ullman 2025; Shapira 2023; Gordon 2016; Le, Boureau, and Nickel 2019; Shapira et al. 2024; Xu et al. 2024; Chen et al. 2024). Social IQa (Sap et al. 2019) targets commonsense reasoning about social interactions, and EmoBench (Sabour et al. 2024) evaluates emotional intelligence. Related work in pragmatics explores figurative language understanding (Liu et al. 2022b, 2022a; Stowe, Utama, and Gurevych 2022; Ma et al. 2025), implicature and presupposition (Zheng et al. 2021; Hu et al. 2023; Jeretic et al. 2020), intention recognition (Sakurai and Miyao 2024), and neologism understanding (Zheng, Ritter, and Xu 2024). MM-SAP (Wang et al. 2024) assesses self-awareness, and Hessel et al. (2023) test understanding of humor.

Recent benchmarking efforts reflect the fundamental shifts in NLP research culture and priorities brought about by the widespread adoption of LLMs. With LLMs becoming increasingly prevalent in daily life, ethical behavior becomes an important dimension of LLM evaluation, such as testing a model's ability to avoid social biases and toxic or harmful outputs (Hartvigsen et al. 2022; Magooda et al. 2023; Zhang et al. 2024c; Liu et al. 2024). Relatedly, as the ability to distinguish between human- and machine-generated text becomes more pressing, benchmarks emerge to assess the models' ability to detect synthetic texts (Sarvazyan et al. 2023; Dugan et al. 2024) and evaluate LLM watermarks (Tu et al. 2024). The increase in possible context window sizes has prompted the development of benchmarks designed to test LLM performance on inputs spanning over a hundred thousand tokens (Zhang et al. 2024b). At the same

time, growing concerns about factual reliability have motivated benchmarks assessing hallucination tendencies (Liang et al. 2024) or robustness (Siska et al. 2024). The increasing popularity of chain-of-thought prompting to elicit intermediate reasoning steps has led to the development of datasets that focus on verifying reasoning chains (Jacovi et al. 2024). Finally, the open-ended nature of LLM generation has increased the importance of format adherence (Jiang et al. 2024b; Xia et al. 2024).

Large-scale and collaborative LLM benchmarking. MMLU (Measuring Massive Multitask Language Understanding, (Hendrycks et al. 2021)) has been the reference benchmark for years to test the abilities of LLMs in commonsense and math reasoning as well as domain knowledge in academic fields like STEM, humanities, social sciences, and law, with the tasks framed as multiple-choice questions adapted from real exams and textbooks. One outstanding problem with LLM benchmarks such as MMLU is that with the rapid progress in LLM development they are quickly *saturating*, i.e., they have become increasingly easy for advanced models to surpass, making it difficult to distinguish subtle differences in model abilities. The development of particularly difficult benchmarks, such as MMLU-Pro (Wang et al. 2024) and Humanity's Last Exam (HLE, Liang et al. (2024)) was aimed at addressing this issue.

In the context of our initiative, the genesis of HLE, which features 2,500 expert-level questions across multiple disciplines, is particularly interesting. To build this resource, the organizers (Scale AI[5], a company which produces curated data and infrastructure to AI companies which develop LLMs) sent out calls through academic communications channels such as faculty mailing lists, graduate programs, and conference communities for subject-matter experts from academia and industry, in multiple fields, such as neuroscience, law, mathematics, medieval history, etc. Eventually, they received contributions from a total of nearly 1,000 contributors across 50 countries and more than 500 institutions. All contributions were voluntary and were structured as fully specified exam-style questions submitted through a secure online platform.

CALAMITA shares with HLE the community-driven spirit which drives development. However, not only does HLE still focus on English, but mostly its creation was motivated by saturating benchmarks, with the aim of creating truly *difficult* questions, where variation is given by the presence of questions in multiple expert fields. Our aim is to create a *diverse* benchmark to test multiple abilities of LLMs in Italian. Lastly, while HLE is a community effort which is company-driven, CALAMITA is a community effort coordinated by the association representing the Italian NLP research community.

The initiative that CALAMITA is closest to is BIG-bench, a community-sourced benchmark, focused on English and launched and coordinated by Google Research, featuring 204 tasks contributed from 442 researchers across 132 institutions (Srivastava et al. 2023). We share with this project attention to diversity and a collaborative, bottom-up strategy to benchmark creation, departing from the top-down approach typical of the more classic benchmarks mentioned above, including MMLU. Yet again, rather than born in the context of a company, CALAMITA is a non-English effort led by the national NLP Association in Italy, with a dynamic structure which enables the continuous addition and evaluation of new tasks (Section 5.2).

Lastly, we would like to mention some very recent efforts by the Iberian NLP community, who in 2024-2025 started two initiatives similar to CALAMITA, namely IberoBench (Baucells et al. 2025) and IberBench (González et al. 2026). In the context of these

5 https://scale.com/

two efforts, researchers mostly took existing benchmarks from previous initiatives such as IberEval and IberLEF (Rosso et al. 2018; Montes-y-Gómez et al. 2023, e.g.), and urged the NLP community to contribute their help to integrate them, correct them, etc. Also due to this strategy, the benchmarks mostly revolve around traditional NLP tasks, such as Sentiment Analysis, Natural Language Inference, Named Entity Recognition. While we also plan to integrate in CALAMITA past datasets (and thus tasks) from previous EVALITA campaigns (coordinating with ongoing efforts (Magnini et al. 2025))[6], with CALAMITA we aimed at building a broad and diverse community first, calling upon a large variety of often not yet tested challenges. While the very recent development of these two Iberian benchmarks provides further evidence that community-driven efforts are both a current need and a powerful strategy in NLP benchmarking, they are separate, and born of independent initiatives. Baucells et al. (2025) explicitly state that "IberBench and IberoBench were developed independently and in parallel, without the involved parties being aware of each other". CALAMITA, instead, is a fully centralized initiative launched and supported by the Italian Association for Computational Linguistics, to minimize dispersion of efforts, maximize harmonization, and truly create a sense of community. The core of this paper lies indeed with the collaborative methodology that we adopted, rather than an assessment and ranking of a wide range of LLMs. The latter effort is something that the CALAMITA benchmark enables as future research.

3. Collaborative Methodology

3.1 Procedure

The CALAMITA organization collective included four chairs (senior academics with an NLP background, all members of the steering board of AILC, who conceived, launched, and coordinated the initiative) and an evaluation team, who took care of data collection and processing, and of model testing. The latter group was formed by three master students, three PhD students, and one postdoc who also served as coordinator of the whole evaluation team. This collective of eleven people represented seven distinct academic affiliations. Ten months elapsed from conception to the first workshop, which featured 22 challenge presentations, published proceedings, and a public leaderboard including most results for two LLMs. We outline below a more fine-grained description of the intermediate steps. Figure 1 illustrates the whole pipeline.

Pre-proposals. By means of a call for pre-proposals distributed through standard communication channels, as well as personal networks, potentially interested parties were informed of the initiative and asked to provide a short proposal with their idea for a challenge that would target specific abilities of LLMs. Proposers were asked to provide a motivation, and as many details as possible on the proposed dataset, with the requirement that it would be natively created in Italian. In the call for challenges, it was specifically mentioned that already existing datasets were also welcome as long as they matched the CALAMITA requirements. The proposers were asked to provide a solid motivation for using an existing dataset within CALAMITA, with the commitment to reframe it for the proposed challenge, should that be necessary. Acceptance was

6 This is already partly happening with the integration of ITAEVAL, which contains some of the past EVALITA tasks, see Section 4.1.22. For a more general discussion of the interaction of CALAMITA and EVALITA see Section 5.2.

Figure 1
CALAMITA pipeline. The figure summarizes the full workflow, from challenge proposal and selection to centralized implementation, model evaluation, validation, reporting, and publication. The rolling process (see Section 5.2) enables the continuous integration of new challenges and models in future cycles.

based on meeting this basic requirement, on the feasibility of dataset creation, where not yet available, and on the conceptual soundness of the proposed challenge. No assessment was based on the practical utility of the task, or on its novelty. All accepted pre-proposals (22 out of 26) were then invited for a full submission to the CALAMITA benchmark, with further instructions (Attanasio et al. 2024a).

Submission. All final participants had to submit three artifacts: 1) a technical report of the data and tasks[7]; 2) the challenge dataset; 3) one GitHub gist for each task describing information required for its technical implementation. The template of the technical report was provided by the CALAMITA team and included sections for the challenge and data description (including source and any copyright and privacy requirements). The submission of the technical reports was handled through the SoftConf platform, where at least two members of the CALAMITA team (from both the chairs and the evaluation team) reviewed the technical report providing suggestions also aimed at increasing coherence across reports. To standardize the access to resources and formatting and promote openness, we recommended participants to share their datasets through HuggingFace[8]. However, in the case of issues related to making the data public (e.g., privacy requirements), it was also possible for participants to submit the data through SoftConf in a dedicated zip file. For the specifics of the GitHub gist, see Section 3.2 and Figure 2. Each challenge could include more than one (sub)task. Each task could have a different metric, if necessary. The participants were also asked, in case of tasks having multiple metrics, to choose the main metric to be used for a given subtask to be used in the final

7 These were eventually collected in the Proceedings of the CLiC-it Conference that the first CALAMITA event was co-located with (Dell'Orletta et al. 2024).
8 https://huggingface.co/

CALAMITA's leaderboard; this is also the metric we use for the aggregated results in Section 4.2.1 and in Table A.1 in the Appendix.

From the beginning of the submission phase, a Slack server was created to manage communication internally and facilitate the distribution of tasks across the evaluation team, but also with all participants. In fact, a dedicated channel was used to support all the challenge proposers who needed help in processing data, identifying the best metrics for their tasks, and creating the gist itself towards a successful submission. Arguably, a submission procedure that requires a descriptive Gist and short pseudo-code was a winning aspect of the collaborative paradigm. Participants did not have to implement the challenges themselves, but rather interact with and monitor our evaluation team's efforts in implementing them in a streamlined, centralized, and standard way. This was particularly helpful for those groups who joined the CALAMITA initiative as non-experts in NLP or data processing in the context of model evaluation.

Gist template

Data ID id of the dataset on HuggingFace
Output Type either 'multiple_choice' or 'open-ended'
Prompt template string of the prompt to be used at inference time
Label Column string for the name of the column in the HuggingFace dataset containing the ground truth
Options list of strings for the possible options, only if output type is 'multiple_choice'
Fewshot details information regarding number of shots to sample and split to sample them from. No additional details were to be supplied for zero-shot
Processing custom pre-processing or post-processing functions to apply
Metrics list of metrics to use for evaluation (either implemented in the *lm-eval-harness* library or defined in the GitHub gist itself)

Figure 2
Gist template.

Evaluation. Each challenge was assigned to one member of the evaluation team, with the task of staying in touch with the proposers, help them with finalizing the data and the metrics when necessary, and then run the evaluation on the Leonardo HPC. The load was spread such that each member implemented at most eight subtasks.

A preliminary leaderboard for LLAMA3.1-8B and ANITA-8B was released to the participants, including all CALAMITA tasks. Participants discussed the process and the obtained results with the member of the evaluation team responsible for their task to flag potential errors and identify necessary fixes. After further improving the codebase for all tasks in need of fixes, a new evaluation run was performed.

Workshop and community sharing. On the occasion of the annual national conference of the Italian Association for Computational Linguistics, CLiC-it 2024[9], CALAMITA received a dedicated slot for presenting the whole framework, the 22 challenges, and the leaderboard. All tasks were presented directly by a representative of the proposing team. A discussion panel followed to devise a route to a sustainable future of CALAMITA.

9 https://clic2024.ilc.cnr.it

Follow up. Two additional models were added to the pool of tested LLMs, namely Minerva-7B-Instruct (Orlando et al. 2024), which was released at the time of the workshop, and LLaMA-3.1-Instruct-70B (Grattafiori et al. 2024) to scale up model complexity. Task organizers were then provided with the outputs from all four models, and were asked to perform error analysis which should be included in a brief one-page report. The 22 resulting reports form the core of the Results section of this paper.

Finally, to ensure the growth of our benchmark, we have just recently initiated a rolling task submission procedure and a call for participation in the evaluation team. Further details are provided in Section 5.2.

3.2 Technical Aspects

While datasets were created by the proposing teams, the submitted data as well as the scoring procedures had to be further processed to enable fully automatic evaluation.

Dataset processing. As mentioned, participants were required either to upload their datasets on Huggingface or to send the datasets' zip file through Softconf, and submit a separate GitHub gist for each subtask. We provide the template used for each GitHub gist in Figure 2.

The evaluation team implemented each task in the `lm-evaluation-harness`[10] (`lm-eval`, for brevity) library (Gao et al. 2024), following the requirements outlined by participants in the GitHub gists. A fork of the library was created to implement the tasks. We chose `lm-eval` because of its focus on reproducibility and the ease with which the library allows for the inclusion of additional tasks. Each sub-task was implemented in a dedicated YAML file, following the formatting expected by the library. This YAML file is used to set up common details in task implementation, like the prompt template or the label column, and was created following the information provided in the GitHub gist by the task proposers.

Following standard `lm-eval` practices, each implemented task fell into one of two categories: Multiple-Choice Question Answering (MCQA) and Open-Ended Generation (OE). For MCQA tasks, we measured accuracy computed by comparing length-normalized log-probabilities—each choice's log-probability is divided by the choice length (unit depends on the task: bytes, tokens, or characters) and the choice with the highest normalized score is picked and used for the aggregate metric defined by the proposers, e.g., F1 Macro. OE tasks required more careful considerations. In some cases, we iterated with the proposers to formalize output parsing, especially for structured outputs such as JSON, or task-specific decoding choices such as the end-of-sentence (EOS) token. Moreover, unless the challenge proposers requested otherwise, we opted not to use chat templates and set the decoding temperature to 0.

Model evaluation. With all tasks implemented in the `lm-eval-harness`, the evaluation runs were performed on LEONARDO (Turisini, Amati, and Cestari 2024), a supercomputer that we could access thanks to the acquisition of an ICRA-C project grant (10K GPU/hours).[11] We tested four different models. A first round of evaluation was run for LLaMAntino-3-ANITA-8B-Inst-DPO-ITA (Polignano, Basile, and Semeraro 2026) and LLaMA-3.1-8B-Instruct (Grattafiori et al. 2024), both sharing the same base

10 `https://github.com/EleutherAI/lm-evaluation-harness`
11 Each node in LEONARDO has 4 NVIDIA A100 64 GB VRAM GPUs.

model (LLAMA3.1-8B); the former is instruction-tuned with instructions (automatically) translated into Italian, while the latter is instruction-tuned with English data. In a second round we also tested Minerva-7B-Instruct (Orlando et al. 2024) and LLaMA-3.1-Instruct-70B (Grattafiori et al. 2024). This allowed us to consider a model trained from scratch with a focus on the Italian language (Minerva-7B-Instruct's pretraining data comprises both Italian and English texts in equal proportions), and a multilingual model with a greater number of parameters to also test the influence of size. Outputs were checked with each task's organizers to ensure they were implemented correctly.

The codebase is publicly available[12], including all CALAMITA (sub)tasks. We also plan to release all of the models' outputs (which are used for the error analyses presented in Section 4) on a public repository alongside this work.

3.3 Hurdles and bottlenecks

As the organizers of BIG-Bench and HLE also note, ensuring consistency and quality across so many datasets and contributors is a very difficult challenge. The centralization of data collection, actual task implementation, and evaluation runs in the hands of a small team of evaluators, plus regular interaction with all the challenge organizers, mitigated this problem. Still, some aspects proved particularly challenging to address, especially when custom implementations were necessary. In most cases, `lm-eval` supported the task implementation required by the authors, but for a few edge cases, some missing functionalities had to be added to `lm-eval`. For example, the library did not support task dependency and chained execution (e.g., Task Y requires the output of Task X), a case occurring in one of the CALAMITA challenges.

The framing of the tasks also presented complexities. While, for example, multiple choice questions were straightforward to implement thanks to the `lm-eval-harness`' built-in log-likelihood based evaluation, standardizing evaluation of more purely generative tasks was not trivial. On the one hand, autoregressive generation is computationally expensive, so that a significantly larger amount of time had to be allocated for evaluation and also debugging. On the other hand, the variability in models' generations, particularly when the difficulty of the tasks or the longer context caused models to struggle, complicated the design of post-processing strategies aimed at extracting the relevant content in the output. This was particularly problematic for tasks that required the models to answer with specific data structures, such as JSON objects. In these cases, it was often necessary to interact with the challenge organizers to establish the limits of acceptable or unacceptable answers. More generally, parsing the models' outputs can lead to significant variations in the final scores.

4. CALAMITA Challenges: Descriptions and Results

The first iteration of CALAMITA resulted in 22 challenges articulated in a total of 95 subtasks, all of which appear on the current leaderboard.[13]

The collaborative spirit at the heart of this initiative has fostered a considerable diversity in the proposed challenges, since researchers from multiple fields are interested in testing different model abilities. To make a better sense of the diverse challenges, we opted for categorizing them into a systematic taxonomy, since it enables a more

12 `https://github.com/calamita-AILC/calamita-eval`
13 Leaderboard accessible at `https://github.com/calamita-AILC/calamita-eval`.

Table 1
Categories of abilities tested by CALAMITA challenges. Challenges (**#Ch**) test general abilities such as knowledge about true facts, commonsense, and logical reasoning or specific NLP-oriented abilities such as code generation or machine translation, and are realized through a number of specific subtasks (**#T**).

Ability tested	Description	#Ch	#T
Commonsense reasoning	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.	10	42
Factual knowledge	Knowledge of concrete, verifiable facts about the world, e.g., definitions, historical events, or scientific concepts.	8	29
Linguistic knowledge	Linguistically motivated tasks that test specific language skills, e.g., word sense disambiguation, coreference resolution, or acceptability judgment.	14	60
Formal reasoning	Ability to understand and use formally logical principles to solve problems, e.g., mathematical problems.	7	22
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.	3	16
Code generation	Ability to generate fully functioning code for a specific programming language.	1	1
Machine translation	Ability to translate a sentence from a source language into another language, with one of the two being Italian.	2	15
Summarization	Ability to create relevant summaries of a given excerpt, e.g., news headline generation or news reduction.	2	4

informative identification of model weaknesses across distinct cognitive and linguistic domains, thereby helping with results and error analysis. Based on observations from previous benchmarking work (see Section 2), we defined eight categories spanning general abilities–commonsense knowledge, factual knowledge, linguistic knowledge, formal reasoning, and fairness and bias–alongside specialized NLP capabilities including code generation, machine translation, and summarization. Table 1 describes each top-level category and reports how many challenges and respective subtasks it includes. Challenges and tasks may belong to more than one category; an overview of challenges with associated ability categories is shown in Table A.3 in the Appendix. In the Appendix, we also report a detailed table with each category and its relative challenges and subtasks (Table A.2).

Under *commonsense knowledge*, we have included tasks such as ECWCA and EurekaRebus, both involving language puzzles calling upon reasoning over general knowledge, INVALSI and MultIT, multiple-choice challenges on general knowledge, and MACID, a challenge revolving around the identification of closely related action concepts. Under *factual knowledge*, in addition to the already mentioned multiple-choice benchmarks, we have included challenges like BEEP, designed to evaluate LLMs on a simulated Italian driver's license exam, and VeryfIT, where LLMs are evaluated on their in-memory factual knowledge on data written by professional fact-checkers. *Linguistic abilities* are of the kind tested in the TRACE-it challenge, where models are tested for

their ability to comprehend a specific type of complex syntactic construction in Italian, in BLM-it which tests the representation and knowledge of abstract linguistic rules in LLMs, or in ABRICOT, where models are tested on their ability to understand and assess the abstractness and inclusiveness of language. The linguistic puzzle challenges are also categorized as *formal reasoning* challenges, together with, for example, GEESE, which assesses the impact of generated explanations on the predictive performance of LLMs in the Textual Entailment task, and GITA, where models are tested for their reasoning capabilities regarding the physical world. Two challenges probe LLMs for their *fairness*, namely GFG, designed to assess and monitor the recognition and generation of gender-fair language in both mono- and cross- lingual scenarios, and PEJORATIVITY, which investigates misogyny expressed through neutral words that can acquire negative connotation when used as pejorative epithets. GFG is also one of the two tasks testing *machine translation* abilities, together with MAGNET, where models are tested in their translation skills, focusing on Italian and English. Lastly, one challenge (TERMITE) focuses on testing *coding* abilities in the context of SQL queries, and one (GATTINA) on the models' summarisation abilities by making them generate headlines for news articles. From the original ITAEVAL benchmark (Attanasio et al. 2024b, 2024c), we have retained only those tasks which had been natively developed in and for Italian, and decided not to include in the official CALAMITA benchmark the datasets derived from translating original English datasets. Nevertheless, through the CALAMITA evaluation framework, we could easily evaluate all of the ITAEVAL tasks, including the translated ones. While these results are not included in any of the aggregated analyses that we present in Section 4.2.1, nor in the overview tables in the Appendix, they are discussed in the paragraph specifically devoted to ITAEVAL results (Section 4.1.22) and reported for completeness in a separate table in the Appendix (Table A.4). Each retained ITAEVAL task was also assigned one or more of our taxonomic categories.[14]

Each challenge, with challenge-specific results and insights, is discussed in Section 4.1. In contrast, Section 4.2 presents a broader discussion with a higher-level analysis across challenges and categories based on aggregated results. Full detailed results per challenge and subtask are included in the Appendix (Table A.1). Please note that details for each challenge regarding background, motivation, data collection and dataset composition, as well as metrics is available in the challenge description papers which have been collected and published as part of the introduction to the CALAMITA initiative. They are collected in Dell'Orletta et al. (2024) and referred to specifically in each of the challenge-specific paragraphs below.

4.1 Challenge-Specific Insights and Lessons Learned

This section reports task-specific insights contributed by the organizers of each challenge after the evaluation campaign.[15] Following the collective run on four open-weight models (LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B, MINERVA-7B), we returned both

14 The ITAEVAL tasks which are not included in CALAMITA due to their being originally developed in and for English are: arc_challenge_ita, belebele_ita, hellaswag_ita, squad_it, truthfulqa_mc2_ita, xcopa_it; see (Attanasio et al. 2024b, 2024c) for more details on these tasks, and Table A.4 in the Appendix for the results on the four models tested within CALAMITA.

15 Each subsection was drafted by the corresponding challenge organizers who received the model outputs from the evaluation team. As a result, prose style and reporting conventions naturally vary across subsections. We have harmonized table formats and model naming; we did our best to uniform figure styles, though minor individual variations still hold. For several challenges Italian examples are shown, but not all of them are translated; in most cases we preserved the choices of the challenge organizers.

scores and raw generations to all teams and asked them to perform a short, focused error analysis. Their one-page reports constitute the backbone of the aggregate analysis which follows, and make this the most collaborative portion of the paper. For each task we provide: (i) a concise task synopsis; (ii) a brief summary of results; and (iii) a critical analysis of typical errors and edge cases, so that aggregate numbers are complemented with linguistic evidence about strengths and weaknesses of each model. This procedure aims to turn a static and rather dry leaderboard into a more actionable understanding of model behavior across heterogeneous tasks, metrics, and prompts.

4.1.1 ABRICOT - ABstRactness and Inclusiveness in COntexT

Puccetti et al. (2024)'s challenge evaluates 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. One of the defining features of human language is its ability to convey both specific and general information using the same lexical item. Consider the following examples: (a)*"The* turtle *escaped from the zoo."* and (b) *"The* turtle *is a reptile."* In (a), *turtle* designates a specific individual, whereas in (b), it generalizes to an entire category of reptiles. This feature of natural languages, called genericity (Krifka et al. 1995) is fundamental to linguistic economy, allowing speakers to efficiently express a wide range of meanings without requiring a separate lexical item for every level of genericity. This task was designed to evaluate the ability of Italian language models to grasp these two semantic properties of word meaning in context: abstractness, the degree to which a noun phrase (NP) refers to a concept rather than a tangible entity, and inclusiveness, the extent to which an NP denotes a broad category encompassing multiple subtypes. The task consists of a dataset of 127 samples, each featuring a specific NP and its corresponding abstractness and inclusiveness scores, as annotated by five human raters on a continuous scale from 0 to 1. Notably, the dataset includes 20 unique NPs that recur across different contexts, allowing for a nuanced assessment of how context influences semantic interpretation. ABRICOT is divided into two subtasks: ABRICOT-ABS, where language models assign a score from 1 to 5 to assess the abstractness of an NP in context. ABRICOT-INC, where models evaluate the NP's inclusiveness using the same scoring scale. Model performance is measured using Pearson correlation with human annotations, quantifying how well computational models approximate human semantic intuition.

Table 2
Results for the ABRICOT task, which evaluates model sensitivity to abstractness (ABS) and inclusiveness (INC) in context. Scores are reported as Pearson correlations with human annotations, comparing four Italian LLMs across the two subtasks.

	LLaMA3.1-8B	LLaMA3.1-70B	Anita-8B	Minerva-7B
ABRICOT–ABS	0.29	0.44	0.50	-0.03
ABRICOT–INC	0.18	0.25	0.14	-0.20

Performance across models. Table 2 shows the Pearson correlation scores for each of the four tested models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B) on the two ABRICOT subtasks. LLAMA3.1-70B outperforms other models on both subtasks, while MINERVA-7B Instruct performs the worst, showing negative correlation for the ABRICOT-INC subtask.

Error Analysis. All tested models are far from human performance. We report an error from the best model, LLAMA3.1-70B: it assigns the same score, 3, to the word *persona* (person) when seen in two contexts: (1) "*Le persone affette da eczema sono spesso sensibili ai prodotti aggressivi*" (en: "*People affected by eczema are often sensitive to aggressive products*") and (2) "*Questa persona è davvero noiosa*" (en: "*This person is really boring*"). However, the first occurrence of *persona* encompasses a large number of individuals, while the second refers to a specific individual.

4.1.2 AMELIA - Argument Mining Evaluation on Legal documents in ItAlian

Grundler et al. (2024)'s challenge consists of three classification tasks in the context of argument mining in the legal domain. The goal of the challenge is to assess the ability of LLMs to understand logical relations in legal reasoning, a particularly challenging field, due to its domain-specific technical language, that uses complex syntactic structures. The tasks are based on a dataset of 225 Italian judicial decisions on Value Added Tax, annotated to identify and categorize argumentative text. The objective of the first task is to classify each argumentative component as a premise or conclusion. The second task aims at predicting the type of a premise as a multi-label classification: legal, factual, or both. A factual premise describes factual situations and events, pertaining to the substance or the procedure of the case, while a legal premise specifies the legal content, such as legal rules, precedents, interpretation of applicable laws and principles. The third task aims at identifying the argumentation scheme, or reasoning patterns, of legal premises in a multi-label classification setting. Given that in all three tasks the classes are unbalanced, the evaluation is based on the macro F1 score.

Table 3
Results for the AMELIA task. For each subtask (Argument Component, Premise Type, and Argument Scheme) we report the F1-score for each class and the macro-average across classes, under zero-shot and few-shot settings.

Model	Shot	Arg Component			Premise Type			Arg Scheme					
		prem	conc	Avg	F	L	Avg	Class	Itpr	Prec	Princ	Rule	Avg
LLAMA3.1-70B	zero	0.93	0.49	0.71	0.86	0.83	0.85	0.18	0.06	**0.76**	0.35	**0.76**	0.42
	few	**0.96**	**0.76**	**0.86**	**0.87**	**0.85**	**0.86**	**0.46**	**0.44**	0.75	**0.42**	0.75	**0.57**
LLAMA3.1-8B	zero	0.83	0.11	0.47	0.43	0.36	0.39	0.12	0.08	0.64	0.23	0.64	0.34
	few	0.93	0.02	0.48	0.76	0.72	0.74	0.22	**0.44**	0.54	0.30	0.59	0.42
ANITA-8B	zero	0.84	0.04	0.44	0.73	0.40	0.57	0.08	0.02	0.18	0.14	0.60	0.21
	few	0.76	0.06	0.41	0.70	0.51	0.60	0.08	0.31	0.24	0.24	0.55	0.28
MINERVA-7B	zero	0.00	0.00	0.00	0.65	0.49	0.57	0.06	0.22	0.21	0.00	0.29	0.16
	few	0.46	0.00	0.23	0.66	0.29	0.48	0.17	0.36	0.42	0.22	0.17	0.27

Performance across models. Table 3 summarises the performance of the four evaluated models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B) across the

three AMELIA subtasks. The best-performing model for all tasks is the largest one, i.e., LLAMA3.1-70B in few-shot mode. It surpasses or matches any other model also in zero-shot mode. It is noteworthy that no other model reaches a macro F1 score above 0.50 in the first and third tasks. The second best model is LLAMA3.1-8B used in few-shot mode.

Error Analysis. Smaller models often do not follow the instructions, generating text similar to few-shot examples, repetition of the expected answers in various formats (e.g., "prem/conc?"), or a non-answer. Focusing on the best-performing model, in the first task, some errors are understandable given the ambiguity of the input. Consider the misclassification of conclusions starting with "*Preliminarmente*" (en: "*Preliminarily*"), or premises that contain expressions like "*in definitiva*" (en: "*Ultimately*"), which seems to indicate the conclusion of an argument. However, further errors involve inputs that are more straightforward, such as the conclusion "*Va infine accolto l'appello incidentale*" (en: "*The cross-appeal must finally be upheld*"). As regards premise type classification, the model often assigns both labels in ambiguous contexts. Another common issue concerns misclassifying normative references as purely factual or legal, especially when abstract norms are applied to concrete scenarios. Consider the premise "*Una statuizione di tal tipo comporta la non opponibilità del provvedimento giurisdizionale* [...]" (en: "*A ruling of this kind entails the non-opposability of the judicial measure [...]*"). The model correctly predicts its legal nature, but fails to capture the factual implication of the legal decision. Concerning the identification of argumentative schemes, the model often misinterprets a normative inference as an interpretative ('Itpr') one. For example, "*Come già evidenziato in primo grado, il fatto che* [...]" (en: "*As already highlighted at first instance, the fact that [...]*"), is a rule-based conclusion and not an interpretative argument. The model also struggles to distinguish between reliance on precedent ('Prec') and the derivation of rules, whenever a decision with normative implications is cited. For example, "*Ed in seguito ai chiarimenti della nota sentenza del 2011*[...]" (en: "*And following the clarifications of the well-known 2011 judgment [...]*"), is incorrectly predicted as both 'Prec' and 'Rule'. The first task, expected to be the easiest, appears to be challenging for most models, especially the identification of conclusions. This is surprising, given that simpler models based on lexical features, such as LinearSVC, obtained better results on similar tasks (Grundler et al. 2022). Alternative prompt formulations might yield improvements regarding the misaligned outputs.

4.1.3 BEEP - BEst DrivEr's License Performer

Mercorio et al. (2024)'s challenge is designed to evaluate LLMs' abilities in a simulated Italian driver's license exam. The task tests both factual knowledge and reasoning skills applied to realistic driving scenarios, requiring understanding of traffic laws, road signs, driving behaviour, and vehicle maintenance. Derived from official ministerial materials, BEEP focuses on the theoretical exam for Category B licenses, which is required for operating cars weighing up to 3.5 tons and with a seating capacity of eight. In Italy, obtaining a driver's license involves strict theoretical and practical assessments regulated by the Ministero delle Infrastrutture e dei Trasporti. The theoretical exam consists of 30 multiple-choice questions, with a maximum of three errors allowed. Beyond rote memorisation, it demands the practical application of rules, particularly important in complex urban traffic contexts. BEEP mirrors this complexity by presenting a series of true/false questions aimed at assessing LLMs' ability to reason about safety, compliance, and real-world driving challenges, aligning with the high standards of the Italian

and European licensing systems. While LLMs have demonstrated strong language understanding, their effectiveness in practical decision-making scenarios, particularly in Italian, remains less explored. Assessing their performance in this structured yet complex setting helps determine their capability to generalise language understanding to real-world applications.

Table 4
Results for the BEEP task. Model accuracy (%) across the main thematic categories of the Italian driver's license exam.

Category	LLAMA3.1-8B	LLAMA3.1-70B	MINERVA-7B	ANITA-8B
DOCUMENTS	54.41%	75.48%	54.02%	51.34%
VEHICLE EQUIPMENT	60.53%	76.32%	51.97%	59.87%
VEHICLES	67.92%	83.96%	56.60%	62.26%
THE MOTOR VEHICLE	64.82%	88.93%	60.08%	59.68%
ACCIDENTS AND INSUR.	69.23%	92.60%	46.30%	67.05%
THE ROAD	58.13%	76.35%	54.19%	57.64%
RULES OF CONDUCT	61.63%	80.04%	50.83%	59.65%
FIRST AID	70.83%	86.46%	54.17%	72.92%
ROAD SIGNAGE	50.00%	75.00%	50.00%	50.00%
SAFETY AND POLLUT.	80.41%	91.84%	51.43%	76.33%
Overall	**64.83%**	**84.25%**	**51.51%**	**62.43%**

Table 5
Results for the BEEP task under the simulated Italian driver's license exam. The table reports the percentage of tests passed (out of 1000 simulations) and the average number of errors per test (with standard deviation).

Model	Total Tests Passed (%)	Avg. Errors (Std.)
LLAMA3.1-8B	5/1000 (0.5%)	10.60 (±2.61)
LLAMA3.1-70B	272/1000 (27.2%)	4.80 (±1.94)
MINERVA-7B	0/1000 (0.0%)	14.64 (±2.74)
ANITA-8B	1/1000 (0.1%)	11.20 (±2.69)

Performance across models. Table 4 presents the Overall Accuracy of the four evaluated models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B). The scaling laws hold as performance improves with an increasing number of parameters. Models achieve higher accuracy in the "SAFETY AND POLLUTION", "FIRST AID," and "ACCIDENTS AND INSURANCE" categories. This could be due to the broader nature of these categories compared to more specialized ones like "DOCUMENTS" or "VEHICLE EQUIPMENT," where performance is lower. We also evaluated the models by simulating an official driving license exam, following official guidelines and introducing another performance indicator. From the dataset, we created 1,000 unique samples of 30 questions each and measured the pass rate, considering a test successful if the model made three or fewer mistakes. As shown in Table 5, smaller models consistently failed,

Table 6
Results for the BLM-IT task. F1 scores of the evaluated models across the three subtasks (Agreement, Causative, Object-drop), averaged over Type-II and Type-III data, under zero-shot and one-shot settings.

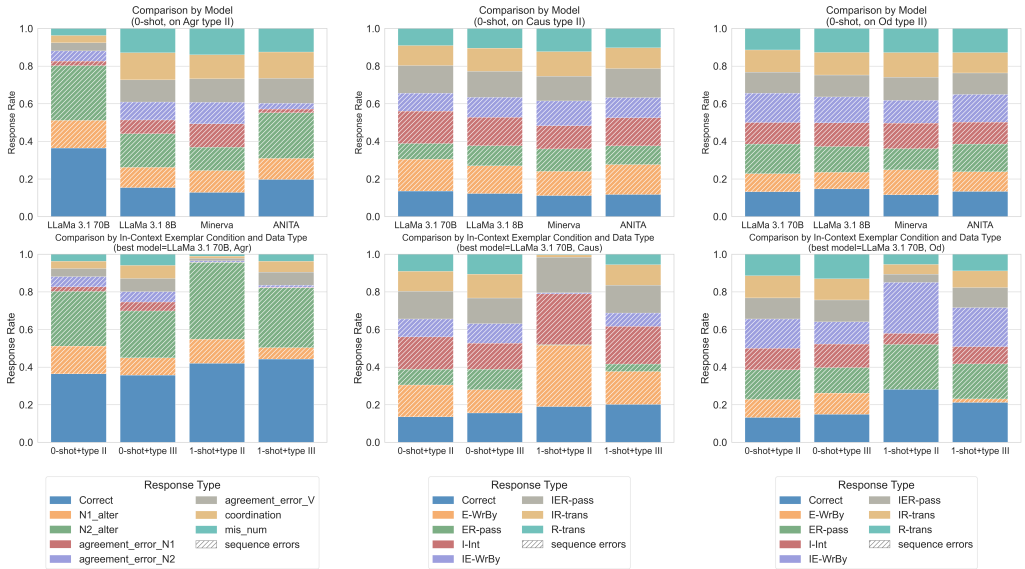
Task	Model	0-shot	1-shot
Agreement	LLAMA3.1-70B	0.339	0.409
	LLAMA3.1-8B	0.104	0.248
	MINERVA-7B	0.026	0.083
	ANITA-8B	0.170	0.242
Causative	LLAMA3.1-70B	0.086	0.186
	LLAMA3.1-8B	0.054	0.131
	MINERVA-7B	0.027	0.090
	ANITA-8B	0.058	0.102
Object-drop	LLAMA3.1-70B	0.071	0.234
	LLAMA3.1-8B	0.063	0.120
	MINERVA-7B	0.028	0.093
	ANITA-8B	0.057	0.094

averaging around 10 errors per test, while even larger models like LLAMA3.1-70B struggled. However, we believe more advanced models, such as GPT-4o, could achieve better results.

4.1.4 BLM-It - Blackbird Language Matrices

Jiang et al. (2024a)'s challenge comprises linguistic puzzles developed in analogy to the visual Raven Progressive Matrices tests, a psychometric test of intelligence. BLMs are developed to test the representation and knowledge of abstract linguistic rules in LLMs (Merlo 2023; An et al. 2023; Samo et al. 2023; Nastase et al. 2024a, 2024b). A BLM instance consists of a context set and an answer set. The context is a sequence of sentences that extensionally encode a linguistic rule. They encode, for example, the rule of grammatical number concord: subject and verb agree in their grammatical number independently of their linear distance. To solve the BLM, one must select, given the context, the missing sentence to complete the sequence. Beside the unique correct answer, the multiple-choice answer set includes minimally contrastive negative examples across different dimensions. The datasets present syntactic and semantic problems in Italian (subject-verb agreement, the causative verb alternation, the object-drop alternation), with a few prompts for a few-shot setting. By their construction, BLM datasets are richly structured at multiple levels: within each sentence, across the input sequence, and within each candidate answer. This structure supports both linguistic and computational investigations, such as testing whether models encode abstractions in their embeddings; linguistic abstraction at the lexical level, or language independent linguistic rules. We evaluate LLM performances on three BLM-It tasks: subject-verb agreement (Agr), causative alternation (Caus), and object-drop alternation (Od), across different dimensions of lexical variation (Type II, Type III; see details in Jiang et al. (2024a)).

Performance across models. Performance results of the four tested models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B) are shown in Table 6, with error dis-



(a) Agreement (b) Causative (c) Object-drop

Figure 3
Response distributions for the BLM-IT task across the three subtasks: (a) Agreement, (b) Causative, and (c) Object-drop.

tributions illustrated in Figure 3. Overall, performance is at or below the baseline level, but several general patterns emerge. Model size has a clear effect, with LLAMA3.1-70B demonstrating the highest performance. Few-shot learning shows varying efficacy across tasks, with the agreement task benefiting the most. Italian-optimized models underperform general-purpose models of comparable scale, with ANITA-8B performing worse than LLAMA3.1-8B in five out of six experimental conditions.

Error Analysis. The most frequent agreement errors relate to the sequence structure of the BLM rather than morpho-syntactic agreement. For verb alternations, the zero-shot setting produces a relatively balanced distribution of errors. In contrast, in the one-shot setting, two prominent sequence errors encode forms of agentivity preference (Huber et al. 2024). Overall, the models master syntax and the syntax-semantics interface, but fail to understand the more general paradigm expressed by the BLM template.

4.1.5 DIMMI - Drug InforMation Mining in Italian

Manna, di Buono, and Giordano (2024)'s challenge aims at evaluating the proficiency of Large Language Models in extracting drug-specific information from Patient Information Leaflets (PILs). The challenge seeks to advance the understanding of effectiveness in processing complex medical information in Italian, and to enhance drug information extraction and pharmacovigilance efforts. The challenge provides a dataset of 600 Italian PILs, derived from the D-LeafIT Corpus[16] (Giordano and di Buono 2024), made up of 1819 Italian drug package leaflets extracted from the Italian Agency for Medications (Agenzia Italiana del Farmaco - AIFA) website[17]. The objective is to test the models'

16 https://github.com/unior-nlp-research-group/D-LeafIT
17 https://www.aifa.gov.it/en/home

ability to accurately answer specific questions related to drug dosage, usage, side effects, drug-drug interactions. For each drug in the dataset, we evaluate the results from two types of zero-shot prompts in Italian: specific task-focused prompts (P1) and structured prompts (P2). P1 is composed of five questions for each of the information types we want to extract. P2 requires the simultaneous extraction of all five categories with a specific instruction to help the model understand the expected output structure and facilitate the extraction process. The answers generated by the models are compared against the gold standard (GS), created to establish a reliable, accurate, and a comprehensive set of answers. For each drug and each information category, the GS contains the correct information extracted from the leaflets through a manual annotation.

Table 7
Results for the DIMMI task. Evaluation under two prompting strategies (P1 and P2). Scores are reported for overall F1 and for each information category: Molecule, Usage, Side effect, Posology, and Drug interaction.

Model	P	F1	Molecule	Usage	Side effect	Posol.	Drug inter.
ANITA-8B	P1	–	0.64	0.12	0.07	0.16	0.21
LLAMA3.1-8B	P1	–	0.86	0.19	0.16	0.17	0.43
LLAMA3.1-70B	P1	–	0.87	0.19	0.15	0.18	0.39
MINERVA-7B	P1	–	0.01	0.08	0.01	0.17	0.12
ANITA-8B	P2	0.33	0.84	0.15	0.08	0.14	0.45
LLAMA3.1-8B	P2	0.31	0.76	0.22	0.05	0.15	0.36
LLAMA3.1-70B	P2	0.34	0.87	0.16	0.06	0.20	0.40
MINERVA-7B	P2	0.07	0.00	0.07	0.00	0.16	0.11

Performance across models. The four models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B) were evaluated under both prompting strategies, P1 and P2 (Table 7), using accuracy, precision, recall and F-1 score against the manually created dataset. LLAMA3.1-70B outperforms all the other models in all categories (in some cases with a very slight difference), except for *Drug_interaction*, where LLAMA3.1-8B performs better in P1 and ANITA-8B in P2. On the other hand, MINERVA-7B consistently failed to extract information for the *Molecule* and *Side_effect* categories in both settings. These two categories differ notably: *Molecule* involves short, specialized terms, making it easier to retrieve, while *Side_effect* includes more varied and often colloquial expressions, increasing extraction difficulty. Prompt type had no effect on MINERVA-7B 's performance, while ANITA-8B improved in most categories with P2. In contrast, LLAMA3.1-8B performed better in P1, except for *Usage*, where it showed a slight improvement in P2 (0.22 vs. 0.18). In P1, LLAMA3.1-8B consistently outperformed ANITA-8B, particularly in *Molecule*, *Side_effect*, and *Drug_Interaction*. However, in P2, ANITA-8B reversed this trend, outperforming LLAMA3.1-8B in those same categories. Both models showed their weakest performance on *Side_effect* in both settings, suggesting intrinsic challenges in that category. Conversely, their best scores were achieved in *Molecule*-Anita in P2, LLAMA3.1-8B in P1. MINERVA-7B reached its peak performance in Posology under both prompts, matching LLAMA3.1-8B 's best score in P1 (0.16). Overall, ANITA-8B benefits from joint-category prompts (P2), while LLAMA3.1-8B performs better with focused, category-specific prompts (P1), except in *Posology*, where

results are comparable. MINERVA-7B 's failure across categories may stem from training limitations or poor compatibility with the task's prompt format.

Error Analysis. A qualitative analysis of selected outputs revealed recurring issues across models. Some responses contained additional, often irrelevant content due to overextended extraction (e.g., ID 147 - Anita - Posology). Others were marked incorrect due to language interference, such as translations or use of synonyms in the wrong language (e.g., ID 556 - Anita - *Molecule*, the model output is irbesartan, hidroclorotiazida instead of the expected irbesartan, idroclorotiazide). Other mismatches arose from inconsistent measurement units (e.g., ID 242 - Anita - *Posology*, the model produces "*80 mg al giorno, 120 mg al giorno*" (en: "*80 mg per day, 120 mg per day*"), whereas the reference answer is "*una compressa al giorno*" (en: "*one tablet per day*")) or failure to capture multiword expressions (e.g., ID 140 - Anita - *Molecule*, the model returns *tamsulosina*, while the correct answer is *tamsulosina cloridrato*). Variations in specificity were also noted, with some outputs being overly generic or, conversely, including unnecessary details. Our analysis also indicates that the models exhibit a tendency to extract layout-based structures, such as bullet points, as undifferentiated blocks rather than parsing the individual, semantically distinct items they contain. This behavior points to a limitation in the models' ability to segment and interpret information accurately when it is presented in visually cohesive formats. In conclusion, these findings expose systematic limitations in model performance, including issues with language handling, term granularity, and expression consistency. They highlight the need for improved prompt design and evaluation metrics, particularly in specialized domains requiring high terminological precision.

4.1.6 ECWCA - Educational CrossWord Clues Answering

Zugarini et al. (2024a)'s challenge 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. Clue-answer pairs were generated following clue-instruct (Zugarini et al. 2024b). Clue-instruct produces three different clues for each given answer and its context. In order to keep at most one clue per answer, a human selection step was added. Examples were constructed from the most visited Italian Wikipedia pages. To guarantee high quality clues, annotators were asked to cross-verify the factual adherence of the chosen clue-answer pairs. Moreover, to better assess factual knowledge of LLMs, annotators were instructed to select clues that provided enough information to infer the correct answer with confidence.

Performance across models. From the evaluation of the four models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B) we see that performances are strongly influenced by two factors: model size and native pre-training. Notably, the character-length hint is generally not beneficial, with mixed results across models. The only model exhibiting distinctly anomalous behavior was ANITA-8B, which often produced outputs with non-standard formatting or irrelevant content, including non-textual characters or unrelated facts. This suggests that native fine-tuning did not improve the factual knowledge of the LLM. On the contrary, native pre-training seems beneficial, since MINERVA-7B exhibits better accuracy than LLAMA3.1-8B, being only slightly inferior to LLAMA3.1-70B, a ten times larger model.

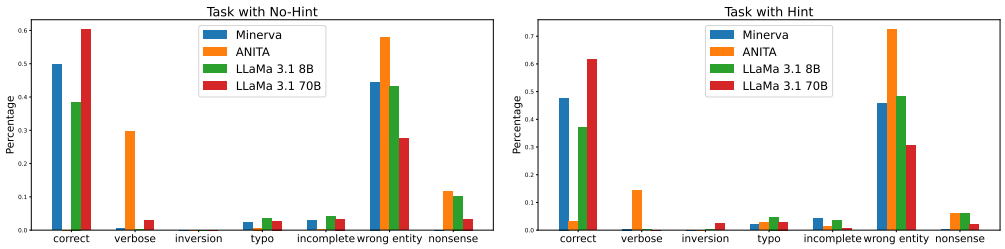


Figure 4
Results for the ECWCA task. Distribution of error types across models in the hint and no-hint settings. Categories include `verbose`, `typo`, `inversion`, `incomplete`, and `wrong entity`. See Table 8 for some error type examples. Correct predictions were marked as `correct`.

Error Analysis. The vast majority of mistakes fall into the `wrong entity` group. Confused entities are often plausible or at least contextually related answers. Sometimes errors stem from queries about events postdating the model's knowledge cut-off, as shown in the example of errors in Table 8. There are also errors from cultural mismatches in non-native models, where the predicted answer reflects a different cultural bias. For example, in the clue "*La sua canzone Gloria è entrata nella classifica dei dischi più venduti in Gran Bretagna e negli Stati Uniti*" (en: "*his/her song Gloria entered the hit of the most sold albums in the UK and in the US*"), where LLAMA3.1-70B predicted "*Laura Branigan*" instead of the Italian singer "*Umberto Tozzi*".

Table 8
Examples of model errors in the ECWCA task, showing representative clues, target answers, predicted answers, and corresponding error types.

Clue	Target Ans.	Predicted Ans.	Error
Comune toscano con importante porto commerciale e turistico	Livorno	Livorno"')	`verbose`
Squalificato per aver morso l'orecchio di Evander Holyfield	Mike Tyson	Tyson Mike	`inversion`
Controparte pasquale del panettone e del pandoro	Colomba Pasquale	Colomba	`incomplete`
Il conduttore di *Lascia o raddoppia?* e *Rischiatutto*	Mike Bongiorno	Mike Buongiorno	`typo`
Trionfatrice al 74-esimo Festival di Sanremo con il brano *La noia*	Angelina Mango	Annalisa Scarrone	`wrong entity`
Attaccante del Napoli e della nazionale belga	Romero Lukaku	Mertens	`wrong entity`

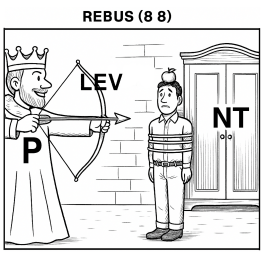
4.1.7 EUREKAREBUS - Verbalized Rebus Solving with LLMs

Sarti et al. (2024a)'s challenge tests the ability of LLMs to conduct multi-step, knowledge-intensive inferences while respecting predefined constraints. The challenge employs text-only variants of *rebus* word games, which have a long-standing tradition in the Italian puzzle-solving community (Tolosani 1901; Miola 2020; Ichino 2021). While turning visuals into textual descriptions simplifies the game, making it akin to cross-

word puzzles, the task remains challenging, requiring multi-step, knowledge-intensive reasoning under predefined formal constraints, such as a specific number of letters. Sarti et al. (2024b) previously showed extremely low accuracy in predicting the final solution (Exact Match) for prompted open-source multilingual LLMs. Since most errors originated from a wrong resolution of definitions, the task was deemed interesting for the CALAMITA campaign to assess the performance of models with an Italian-focused training such as ANITA-8B and MINERVA-7B. In EUREKAREBUS, 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.

Table 9
Results for the EUREKAREBUS task. Evaluation of model accuracy (**Acc.**), predicted word length (**Len.**), and exact match (**EM**) for first-pass and final solution predictions, under prompts without (top) and with (bottom) length hints. The accuracy values reported in Appendix A.1 correspond to the First-Pass Accuracy (FP Acc.) in the two settings.

Model	First Pass (FP)			Solution		
	Acc.	Len.	EM	Acc.	Len.	EM
Without length hints						
LLAMA3.1-8B	0.10	0.18	0.00	0.02	0.17	0.00
LLAMA3.1-70B	**0.36**	0.42	**0.07**	**0.08**	**0.26**	0.00
ANITA-8B	0.00	0.01	0.00	0.00	0.00	0.00
MINERVA-7B	0.00	0.00	0.00	0.00	0.01	0.00
With length hints						
LLAMA3.1-8B	0.07	0.16	0.00	0.01	0.09	0.00
LLAMA3.1-70B	0.32	**0.49**	0.04	0.06	0.21	**0.01**
ANITA-8B	0.00	0.01	0.00	0.00	0.00	0.00
MINERVA-7B	0.00	0.00	0.00	0.00	0.00	0.00



Example of multimodal rebus. FP: P *re* LEV *arco* NT *ante*. Solution: *Prelevar contante*.

Performance across models. Table 9 reports metric scores for the four models (LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B and MINERVA-7B) across the two tested settings. Results confirm that the task remains unsolved by current open-source models, with the best solution accuracy (Exact Match) of 1%. Italian-trained models are also found to struggle with all the task phases compared to the multilingual LLaMA model of comparable size. Major gains are obtained by increasing the parameter count to 70B, confirming the importance of model scale in complex reasoning tasks. LLAMA3.1-70B is the only tested model using length hints in the prompt to resolve FP words (FP Len. 0.49 vs. 0.42 without hints), although this does not lead to better definition resolution. Still, all tested models struggle with respecting the given reasoning template.

Error Analysis. As shown by Sarti et al. (2024b), models can disregard resolved words to generate potentially more fluent solutions. In the following example, LLAMA3.1-70B correctly resolves the first pass, but hallucinates a final solution (optional length hints in blue, Table 9 shows a multimodal version):

Rebus: P – *Il vertice della nobiltà* (2) – LEV – *Quello di Trionfo si trova a Parigi* (4) – NT – *Portiere d' armadio* (4)
Solution key: 8 8
First pass: P *re* LEV *arco* NT *ante* (en: P *king* LEV *arch* NT *doors*)
Model prediction: Per levar antenne (en: For removing antennas)
Correct solution: Prelevar contante (en: Withdraw cash)

Our results confirm that word games can be a promising testbed for evaluating the knowledge and cultural understanding of LLMs while also testing their sequential instruction following skills. Future work in this area could expand our evaluation to multimodal puzzles and less structured LM reasoning templates (Zelikman et al. 2024).

4.1.8 GATTINA - GenerAtion of TiTles for Italian News Articles

Francis et al. (2024)'s challenge aims to assess the ability of LLMs to generate headlines for scientific news articles. Apart from testing the models' performance in a real-life use case, this task requires several foundational skills. First, the model must concisely summarize the content of an article which may be rather long. Second, the generated headline should be attractive and should encourage interaction from readers. Third, the model should adapt to the tone of the article itself, while adhering to the tone of the overall journal.

GATTINA consists of 30,461 article-headline pairs, all of which are human-authored, from two prominent Italian media outlets, 'Ansa scienza' and 'Galileo'. Due to the subjective and context-sensitive nature of headline evaluation, articles are evaluated along several dimensions. For this purpose, two classifiers are introduced: one judges whether a generated headline 'belongs to' a given article, and one assesses whether an article is machine- or human-authored. The former is trained to discriminate between coherent and incoherent headline-article pairs. The latter assesses how close to human writing the generated headline is (though machine-generated, being classified as human-produced is a positive outcome). We additionally use ROUGE (Lin 2004) and SBERT (Reimers and Gurevych 2019) to evaluate content appropriateness.

Table 10
Results for the GATTINA task. Automatic evaluation scores reported for Natural-Synthetic (N-S) and Headline-Article (H-A) metrics across all evaluated models.

Model	N–S	H–A
LLAMA3.1-8B	0.066	0.683
MINERVA-7B	0.138	0.654
ANITA-8B	0.201	0.615
LLAMA3.1-70B	0.441	0.686

Performance across models. With the GATTINA challenge, which contains around 30,000 article-title pairs, LLMs are tested to generate appropriate headlines from news articles. For the same reason, evaluation is far from trivial. To quantify performance, we trained two classifiers: "N-S" (discriminating between a Natural (human-written) and a Synthetic headline, and "H-A" (capturing the relevance of the Headline for the given Article.) Table 10 reports these metrics for the tested models.[18] While this is a

18 Further details on the training of these metrics are provided in (Francis et al. 2024).

true generative task that aligns with the training objective of language models better than many of the other CALAMITA challenges (where classification tasks are recast as generative), the task itself turned out to be far from simple for all of the tested models. This is partly due to the specifics of news headlines, which encapsulate human-intuitive characteristics, such as attractiveness, which are hard to grasp by machines (Cafagna et al. 2019), and partly ascribable to the appropriateness of evaluation metrics.

Error Analysis. Manual evaluation was performed by two professional journalists on a small sample of generated headlines. Overall, the LLMs show limited sensitivity to character count constraints (*Registrato il sussurro dei piccoli delle megattere* [Ansa]=> "*Il sussurro dei piccoli, un segreto per sopravvivere: le megattere 'tacciono' per non farsi sentire*" [ANITA-8B]) – a strict requirement for many news outlets – and tend to favour definitive phrasing, which leaves less room for ambiguity compared to headlines produced by editorial teams. Syntactically, headlines generated by LLMs often use the pattern "key phrase: [text]", such as "*La Luna si restringe: terremoti e frane minacciano le future missioni umane*" (LLAMA3.1-8B). While this is theoretically a correct strategy, it is less favoured by editorial teams, who tend to prefer more fluent phrasing, integrating the keyword within the body of the headline ("*La Luna si sta restringendo, ha perso 45 metri in milioni di anni*" (ANSA).) Models definitely overuse this strategy: 40% of the cases in LLAMA3.1-70B and ANITA-8B, and around 90% in MINERVA-7B and LLAMA3.1-8B. Lastly, LLMs tend to include a greater number of details in the headlines, whereas editorial practices typically emphasise a single aspect of the news story which is expected to most attract the readers' attention.

In addition to stylistic issues, factual subtle aspects were also identified as problematic in generated titles. For example, the generated title "*L'allunaggio giapponese: un successo a metà*" (MINERVA-7B) suggests a half-success of the lunar mission, which is not true to the contents of the article, reflected of course in the original headline ("*Luna difficile, arrivo con batticuore del lander Slim*" (ANSA)). Similarly, "*Il futuro è già qui, ma l'Italia rischia di restare indietro*" (LLAMA3.1-70B) underscores a negative perspective which is the opposite of the article's contents and the original title ("*Biotecnologie, in Italia confermata la crescita a 2 cifre*" (ANSA)). The experts also spotted cases where the generated title is dishonest and click-baiting: "*La mappa nascosta della pandemia: scoperta la vera causa della diffusione del coronavirus*" (LLAMA3.1-8B) (original: "*Pandemia, gli eventi di superdiffusione hanno un ruolo chiave*" (Galileo)): the use of "*map*" is misleading and incorrect, and also the reference to the "*real*" cause of Coronavirus is a potentially dangerous statement, not present in the article.

In some cases the journalists assessed the generated headlines as better than the original ones, such as the title "*Terremoto in Turchia, le onde sismiche arrivano in Antartide*" generated by Anita vs the human-produced (ANSA) title "*Sisma Turchia-Siria, rilevato da sismografi in Antartide*", which includes the repetition of "*sisma*" (*earthquake*) and "*sismografi*" (*seismographs*). Another headline evaluated as better than the original one is "*La tecnica Crispr per il taglia-incolla del Dna potrebbe aumentare il rischio di tumore*" (LLAMA3.1-70B) vs the original (ANSA) "*Forse rischi di tumore dal taglia-incolla del Dna*". Lastly, it appears that models also introduce some degree of creativity, using figurative language and compact titles, such as in the headline "*Il viaggio delle lame, un filo che lega il passato*" (ANITA-8B) vs. the original from Galileo "*9000 anni fa l'ossidiana viaggiava in slitta verso l'Artico*". The expert evaluation here is uncertain: while the title seems good at first glance, especially because of the choice of "*lama*" (*blade*) over "*ossidiana*" (*obsidian*), a word definitely more attractive for a title, the figurative expression "*un filo che lega il*

passato" (*a wire connecting the past*), while suggesting a good semantic connection with "*blades*", remains possibly too cryptic to serve as a safe, valid title.

Dealing with the long context of entire media outlets' articles, being coherent and, most of all, being able to forge a creative headline like a professional journalist still stands as a hard challenge even for the bigger models. In the context of CALAMITA, the title being in Italian instead of English possibly further increases the difficulty of this task. While none of the current models would be able to directly substitute a professional writer in yielding the most appropriate headline given an article, the GATTINA challenge shows that LLMs, especially larger ones, have great potential as part of a journalist's toolbox.

4.1.9 GEESE - Generating and Evaluating Explanations for Semantic Entailment

Zaninello and Magnini (2024)'s challenge 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, different 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. In GEESE, an LLM (M_2) is asked to predict the entailment relation (*entailment*, *neutral* or *contradiction*) between pairs of sentences from the e-RTE-3-it dataset (Zaninello, Brenna, and Magnini 2023). Explanations either written by humans (`gold`) or generated by another LLM (M_1), which may or may not be the same as M_2, are used as a "hint" by M_2. The results are compared with two baselines: M_2 using an empty string (`no-exp`) or M_2 using a copy of the first sentence (`dummy`) as hint. The task assesses M_2 models' ability to profit from explanations (by comparison with the baselines) but also, by comparing the results across different explanations on a given M_2, gives us an indirect measure of the effectiveness of such explanations, which can be taken as a proxy of their quality. Results for the four tested M_2 models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B) are reported in Table 11.

Table 11
Results for the GEESE task. Accuracy of M_2 models on the e-RTE-3-it dataset under different explanation settings: baselines (*no-exp*, *dummy*), human-written (*gold*), and model-generated explanations. The best-performing M_1 model for each M_2 is shown in bold.

	LLAMA3.1-8B M_2	**LLAMA3.1-70B** M_2	**MINERVA-7B** M_2	**ANITA-8B** M_2
No-exp (b.l.)	0.4950	**0.5650**	0.4875	0.4675
Dummy (b.l.)	0.4937	0.6062	0.4850	0.5350
Gold	0.6162	0.8150	**0.5762**	0.7062
LLAMA3.1-8B (M_1)	0.6025	0.8637	0.5662	0.7850
ANITA-8B (M_1)	**0.6625**	**0.8850**	0.5650	**0.8287**

Performance across models. The biggest LLAMA3.1-70B model achieves the highest result of 88.50%. Models utilizing machine-generated explanations generally achieve highest accuracy, with only one model (MINERVA-7B) scoring highest by using human-written explanations (57.62%). Explanations generated by ANITA-8B as M_1 are the most effective in enhancing the predictive ability of M_2's, again with the exception of

MINERVA-7B which, after the human-written ones, favours explanations generated by LLAMA3.1-8B by a small margin (56.62%). In terms of accuracy gain, ANITA-8B as M_2 shows the highest increase of 36% and 29% over the `no-exp` and `dummy` baselines, respectively.

Error Analysis. These results suggest that the quality and type of explanations influence predictive performance, highlighting the need for further research into optimizing explanation generation. For example, in the *no-exp* setting, M_2-LLAMA3.1-8B incorrectly predicts "entailment" for the sentences with id=7, and using the explanations generated by itself (M_1-LLAMA3.1-8B) does change the label, but to "contradiction", which is also wrong. However, using the `gold` and ANITA-8B explanations leads to a correct "neutral" prediction: (LLAMA3.1-8B) *La premessa e l'ipotesi non sono correlate in maniera stretta, poiché la premessa parla della partenza della signora Minton da un paese, mentre l'ipotesi afferma la sua nascita in quel paese. Non vi è un nesso logico tra i due fatti.* (gold) *Il fatto che la signora Minton fosse in Australia prima di andare a Londra non necessariamente significa che sia nata in Australia.* (ANITA-8B): *La premessa afferma che la signora Minton lasciò l'Australia nel 1961, ma non dice nulla sulla sua nascita. L'ipotesi afferma che la signora Minton è nata in Australia, ma non dice nulla sul suo viaggio in Inghilterra. Queste due affermazioni non si riferiscono direttamente l'una all'altra, quindi non c'è alcun collegamento logico tra di esse.*. This is somehow in contrast with the expectation that a model should be better guided by explanations produced by itself. These findings encourage us to further explore the interplay between explanations and model performance, paving the way for more interpretable, transparent and user-friendly AI systems.

4.1.10 GFG - Gender-Fair Generation

Frenda et al. (2024)'s challenge is 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.

Table 12
Results for the GFG task across all subtasks. Evaluation metrics include BScF1 (BERTScore F1), AccG (Accuracy Gente), and CWA (Coverage-Weighted Accuracy).

	Task 1		Task 2			Task 3	
Dataset	**NeoGate**	**GFL**	**GFL**	**NeoGate**	**Gente**	**NeoGate**	**Gente**
Metric	**BScF1**	**BScF1**	**AccG**	**CWA**	**AccG**	**CWA**	**AccG**
ANITA-8B	0.55	0.45	0.51	0.54	0.50	0.35	0.57
LLAMA3.1-8B	0.66	0.50	0.18	0.53	0.21	0.42	0.61
LLAMA3.1-70B	0.59	0.53	0.61	0.73	0.54	0.58	0.56
MINERVA-7B	0.49	0.17	0.53	0.45	0.33	0.28	0.59

Performance across models. Table 12 reports the performance of all four models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B) across all sub-tasks. In general, we observe that the performance of models depends on the task, without an overall tendency. In some cases the smaller LLAMA3.1-8B even outperforms LLAMA3.1-70B.

Error Analysis. We conducted a preliminary qualitative analysis by manually reviewing examples of model outputs for each task. Such initial investigation revealed that the overall quality of the generated outputs was quite poor. Therefore, we decided to conduct a more formal analysis focusing specifically on the model showing good results across all the tasks: LLAMA3.1-70B. We analyzed 50 randomly selected outputs for each task and dataset combination. Each example was manually annotated taking into account several task-related aspects, such as whether the generated text is a direct copy of the original, the presence of hallucinations or meaningless language, and if none, some or all of expressions referring to human entities are gender-neutral in the generated text. We further took note of what strategies the model used (e.g., periphrases employing '*persona*'$_{\text{[person]}}$), whether such strategy was part of the prompt and the model simply reproduced it, and whether the rewriting was less informative than the original text (e.g., if it turned "*ricercatori*"$_{\text{[researchers]}}$, into "*persone che lavorano*"$_{\text{[people who work]}}$).

In the gendered language detection task, the model correctly identified all target spans in 45% of test sentences and none in 20%. It was more prone to selecting unrelated spans in GFL-it than in Neo-GATE (38% vs <3%), possibly due to differences in textual domains: Neo-GATE features more generic language, while the administrative style of GFL-it makes gendered language harder to detect.

Looking at the inclusive reformulation task, the manual analysis confirms the behavior suggested by the metrics: the model performs better in using *innovative* obscuration strategies rather than the *conservative* ones. Among the Neo-GATE outputs, 64% replaced gender markers with neomorphemes, while only 6% preserved them. In contrast, only 40% of GFL-it and GeNTE outputs were fully gender-neutral, and 38% remained fully gendered. However, neomorphemes were often misused: in 29% of outputs, they appeared in unrelated or already neutral words (e.g., '*lə opzione*'$_{\text{[the option]}}$, pazientə$_{\text{[patient]}}$), and in 28%, they were used in required contexts but with incorrect word forms (e.g., '*delə*' instead of '*dellə*'$_{\text{[of the]}}$). This suggests that, despite high CWA, the model struggles with precise neomorpheme generation. In using conservative strategies, the model consistently uses the term '*persona*'$_{\text{[person]}}$ to rewrite generic masculine formulations. Moreover, in 26.5% of the outputs the model simply copied the original text without modifications. In 21.43% of cases information was lost in the rewritten sentence and 61% of the re-writings were included in the prompt. In few cases the model used the double form by adding the female variant, a *visibility* strategy (Rosola et al. 2023) we consider incorrect within this challenge.

Finally, in the inclusive translation task we observe similar trends, though with lower accuracy and less precise use of obscuration strategies, likely due to the added complexity of cross-lingual translation. Among the analyzed outputs, 46% of Neo-GATE translations were fully gender-inclusive and 26% fully gendered, while only 6% of GeNTE outputs were inclusive and 82% gendered. Compared to reformulation, neomorphemes were used less frequently and misused more often: 40% appeared in unrelated or already neutral words. However, incorrect or unnecessary neomorpheme generation was less frequent, occurring in 10% and 16% of outputs, respectively. Notably, among the few inclusive GeNTE outputs, half used formulations with '*persona*.'

As previous assessments of the topic suggest (Savoldi et al. 2023; Piergentili et al. 2024; Piazzolla, Savoldi, and Bentivogli 2023), even modern models that are the state-of-the-art in other tasks struggle with the evaluation and generation of inclusive language. Our analysis confirms that smaller models perform poorly on generative tasks, while larger models occasionally produce inclusive outputs but often employ the same rewriting strategy (e.g., consistently adding '*persona*'). The performance of the LLMs tested shows that there is still room for improvement in the various subtasks proposed and their evaluation.

4.1.11 GITA4CALAMITA - Graded Italian Annotated Dataset

Pensa et al. (2024)'s challenge investigates the reasoning abilities of LLMs in physical commonsense reasoning and introduces a methodology to assess their understanding of the physical world through the design and creation of the Graded Italian Annotated dataset (GITA). GITA is written and annotated by a professional linguist and is composed of plausible and implausible stories. There are two mechanisms to generate implausible stories: cloze implausible stories are created by altering one sentence in a plausible story to make it implausible; order implausible stories are created by switching the order of two sentences in one plausible story to make it implausible. Our benchmark aims to evaluate three distinct levels of commonsense understanding with three specific tasks: 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. Our findings reveal that, although the models may excel at high-level classification tasks, their reasoning is inconsistent and unverifiable, as they fail to capture intermediate evidence.

Performance across models. The performance of the four evaluated models (LLAMA3.1-8B, LLAMA3.1-70B, MINERVA-7B and ANITA-8B), shown in Table 13, is assessed on three subtasks: story classification, conflict detection, and physical state recognition (corresponding metrics: accuracy, consistency, and verifiability). LLAMA3.1-70B demonstrates the best overall performance, achieving 87.64% accuracy in story classification and excelling in conflict detection and physical state recognition. LLAMA3.1-8B performs well, particularly on the cloze dataset, with 72.47% accuracy and notable consistency in conflict detection. ANITA-8B outperforms MINERVA-7B, which struggles significantly, achieving only 38.20% accuracy in story classification and failing in conflict detection and physical state recognition. Overall, the models perform better on the cloze dataset than on the order dataset.

Table 13
Results for the GITA task. The table reports model performance across the three evaluation dimensions- *story classification accuracy*, *conflict consistency*, and *physical-state verifiability*- each computed for the overall dataset as well as for the cloze, order, and plausibility subsets.

Model	Accuracy				Consistency			Verifiability		
	Over.	Cloze	Order	Plaus.	Over.	Cloze	Ord.	Over.	Cloze	Ord.
MINERVA-7B	38.20	13.67	13.11	88.88	1.67	3.41	0.00	0.00	0.00	0.00
ANITA-8B	58.98	61.53	38.52	77.77	18.41	29.91	7.37	7.94	14.52	1.63
LLAMA3.1-8B	72.47	81.19	64.75	71.79	29.28	44.44	14.75	14.22	23.93	4.91
LLAMA3.1-70B	87.64	95.72	88.52	78.63	65.27	74.35	55.73	36.40	47.00	25.40

Error Analysis. Regarding the physical state recognition, models struggle to predict the verifiability of a story (deepest subtask). In particular, MINERVA-7B fails completely with an overall verifiability of 0.0. Oppositely, the Llama models reach the highest scores, with LLAMA3.1-70B outperforming the rest. LLAMA3.1-70B correctly predicts 86 physical states out of a total of 239 implausible stories, corresponding to 12 out of 14 physical state labels, including *occupied*, *solid*, *wet*, and *power*. It excels in detecting *power* states (17 out of 24). In the following story, LLAMA3.1-70B recognizes the impossibility of cooking broth on a turned-off stove: *Marta ha acceso il gas. Marta ha messo sui fornelli una pentola. Marta ha versato il brodo nella pentola. Marta ha spento il gas. Marta ha cucinato il brodo per un'ora.* LLAMA3.1-8B detects a total of 34 correct physical states in implausible stories, with 7 predicted physical state labels out of 14. States related to *functional*, *temperature*, and *in pieces* account for 28%, 20% and 29%, respectively, of the total physical states in the entire implausible dataset. In the following example: *Claudio accende la TV. La TV è rotta. Claudio trova il suo film preferito. Claudio guarda il film. Claudio spegne la TV e va a letto.*, LLAMA3.1-8B identifies the *functional* physical state as the cause of conflict: if the television is not working, the actor of the story cannot watch a movie. Although the LLaMA models achieve the highest scores for this task, there is a noticeable discrepancy between the physical states they predicted. The physical states that are accurately predicted by one LLaMA model are not recognized by the other. We highlight that the cloze dataset outperforms the order dataset in all subtasks. Unlike the cloze set, where conflicts are created by substituting sentences, the order dataset generates conflicts by inverting the order of sentences without adding different words or explicit negations. To conclude, we see that while some LLMs perform well in commonsense reasoning tasks at the end-task level, they lack a robust reasoning process when a deeper understanding of the physical world is required.

4.1.12 INVALSI

Puccetti, Cassese, and Esuli (2024)'s challenge is based on the Invalsi tests administered to students within the Italian school system. The Invalsi tests are country-wide assessments designed by expert pedagogists to monitor the average performance of students over the years, administered multiple times from primary school through high school.[19] The results of these tests have been used in several population studies (Bolondi and Cascella 2017; Costanzo and Desimoni 2017; Pietschnig et al. 2023), but they have seen little use as a benchmark for Large Language Models' understanding and reasoning in Italian. The Invalsi Task is composed of two subtasks, Invalsi ITA and Invalsi MATE, the first focused on language understanding and the second on math reasoning. The Invalsi ITA dataset used in this challenge contains 1117 questions of two types: (1) multiple choice: the student is asked to pick the correct answer among four candidate answers; (2) binary: the student is asked to assess a binary property of a statement, e.g. True – False, Before – After, etc. The Invalsi MATE dataset is composed of 400 questions of three types: (1) multiple choice: the student is asked to pick the correct answer among four candidate answers; (2) true/false: the student is asked to assess whether a given statement is True or False; (3) number: the student is asked the correct numerical answer to a question. In CALAMITA, all questions in both datasets are framed as multiple choice questions and we use a likelihood based approach to evaluate the models, we pick the answer with higher likelihood according to the model. For more details about the benchmark we refer to the work from Puccetti, Cassese, and Esuli (2024).

19 https://www.invalsi.it/invalsi/index.php

Table 14
Results for the INVALSI task, reported separately for the Italian language understanding subtask (ITA) and the mathematics reasoning subtask (MATE). Scores represent model accuracy across all questions.

Task	LLAMA3.1-8B	LLAMA3.1-70B	ANITA-8B	MINERVA-7B
ITA	0.71	0.89	0.71	0.38
MATE	0.51	0.72	0.47	0.34

Performance across models. We report the results for the four models (LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B and MINERVA-7B) in Table 14. We see that model size is more important than the training language, as LLAMA3.1-70B is the only model performing significantly better than the others, including those with Italian tuning.

Error Analysis. The performance gap between Invalsi ITA and Invalsi MATE is explained by knowing that Invalsi ITA questions often require to look for information in a longer text fragment while Invalsi MATE requires to perform a moderate reasoning step. It is challenging to identify question types that are more challenging for LLMs. A trend, reported by Puccetti, Cassese, and Esuli (2025), the Invalsi ITA questions show a clear pattern where difficulty increases as students' age increases, while for Invalsi MATE this is only evident between first grade students and older ones.

4.1.13 ITA-SENSE - ITAlian word SENSE disambiguation

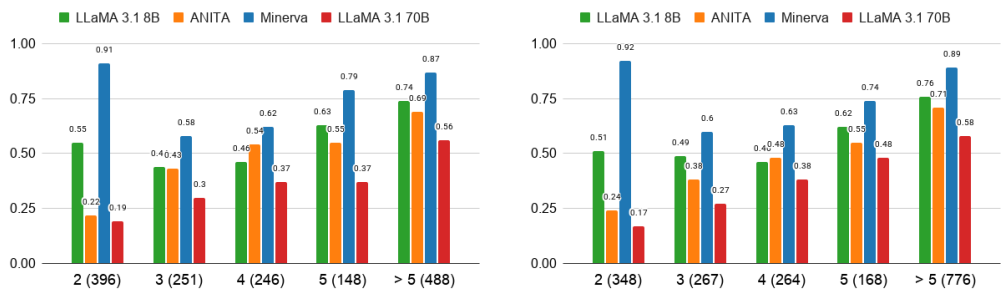
Basile, Musacchio, and Siciliani (2024)'s challenge assesses the abilities of LLMs in understanding lexical semantics through Word Sense Disambiguation (WSD). The classical WSD task is cast as a generative problem formalized as two tasks: [T1] Given a target word and a sentence in which the word occurs, 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. BabelNet (Navigli and Ponzetto 2010) is used as a sense inventory for retrieving the glosses assigned to each synset. Since some synsets have no Italian glosses, we obtained them by translating from English glosses. We provide two test sets: 1) *Original* contains only synsets for which Italian glosses are available; 2) *Translated*, which also contains translated glosses. The four CALAMITA LLMs (LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B and MINERVA-7B) are tested in a zero-shot setting.

Performance across models. We report the performance obtained for each type of task of ITA-SENSE (Basile, Musacchio, and Siciliani 2024) in Table 15. Overall, the impact of using translated glosses is more evident in the multiple choice task, with a significant drop in accuracy. We believe this is caused by the increased number of options (more options are available since there are more glosses).

Error Analysis. We want to understand which words the models found more difficult to disambiguate. In particular, we want to understand whether words with more possible meanings have a more significant impact on the overall error rate. In Figure 5, we plot the error rates for the multiple choice task for four different levels of polysemy. In most cases, we find a higher error rate when there are more possible meanings for the target word. There are some exceptions to this pattern, notably MINERVA-7B and LLAMA3.1-

Table 15
Results for the ITA-SENSE task across both subtasks: [T1] generative and [T2] multiple-choice. R-BERT denotes the harmonic mean of Rouge-L and BERTScore for the generative task, while Acc. indicates accuracy for the multiple-choice task. Results are reported for test sets with and without translated glosses.

	LLAMA3.1-8B		ANITA-8B		MINERVA-7B		LLAMA3.1-70B	
Test set	**R-BERT**	**Acc.**	**R-BERT**	**Acc.**	**R-BERT**	**Acc.**	**R-BERT**	**Acc.**
Original	0.319	0.413	0.259	0.512	0.257	0.215	0.314	0.629
Translated	0.319	0.386	0.263	0.475	0.255	0.200	0.310	0.579



(a) Without Translated Glosses

(b) With Translated Glosses

Figure 5
Error rates for the ITA-SENSE multiple-choice subtask across different levels of word polysemy. Each bar corresponds to words with two, three, or more meanings; values in parentheses indicate the number of instances per group.

8B without translated glosses. Finally, for the generative task, we note that due to the harmonic mean of Rouge-L and BERTScore, instances with no syntactic overlap were assigned a score of 0. However, this is not ideal, as the model can still generate a meaningful and correct definition without using any words from the ground truth. Considering only the BERTscore, we obtain a value of around 0.69 for all LLAMA3.1-8B models and 0.66 for MINERVA-7B. To overcome these issues, we plan to change the metric for evaluating the quality of the generated gloss.

4.1.14 MACID - Multimodal ACtion IDentification challenge

Ravelli, Varvara, and Gregori (2024)'s challenge aims at evaluating LLMs' abilities 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. In each set, the sentences may contain the exact same verb or up to four different verbs, thus, the challenge comprises various levels of difficulty. The dataset is composed of 100 quadruples of sentences derived from the LSMDC Dataset (Large Scale Movie Description Challenge, Rohrbach et al. (2017)), though readapted to Italian after manual translation, and each caption is annotated with a concept from the IMAGACT ontology of actions (Moneglia et al. 2014). Although currently monomodal (text-only), the task is designed for future multimodal integration thanks to the alignment with videos

from LSMDC, linking visual and textual representations to enhance action recognition. The dataset focuses on "pushing" events, meaning all sentences refer to an action that can be described with a verb related to the semantic field of "push". It presents action-predicate mismatches, where the same verb may describe different actions (e.g. "pressing a button" and "pressing the wood") or different verbs may refer to the same action (e.g. "pressing a button" and "pushing a button"). Although currently unimodal (text-only), the dataset aligns each sentence with a short video from the LSMDC dataset (Rohrbach et al. 2017), thus being designed for future multimodal integration.

Performance across models. Overall, the accuracy among the four tested models ranges between 0.27 (MINERVA-7B) and 0.58 (LLAMA3.1-70B), with intermediate values for LLAMA3.1-8B (0.43) and ANITA-8B (0.43). In general, the task is challenging for LLMs, with differences based on size (the larger, the better).

Table 16
Results for the MACID task, reporting model accuracy across different verb distributions within each sentence quadruple. Categories indicate the number and arrangement of distinct verbs: 4 = four different verbs; 3 = two shared + two unique verbs; 2(2+2) = two verbs equally distributed; 2(3+1) shared = the intruder shares its verb with two others; 2(3+1) unique = the intruder has a unique verb; 1 = all sentences share the same verb. Scores represent average accuracy per group and overall.

Verbs dist.	Items	LLAMA3.1-8B	LLAMA3.1-70B	MINERVA-7B	ANITA-8B
4	7	0.57	0.86	0.14	0.57
3	16	0.44	0.50	0.25	0.44
2 (2+2)	9	**0.67**	0.56	0.33	0.44
2 (3+1) shared	23	0.22	0.22	0.30	0.30
2 (3+1) unique	21	**0.67**	**0.90**	**0.43**	**0.62**
1	24	0.29	0.63	0.13	0.33
Overall	100	0.43	0.58	0.27	0.43

Error Analysis. We investigate whether models rely more on the lexical component, i.e. verb lexemes, rather than on concept similarity. Indeed, accuracy is higher when the intruder action is lexically discriminated by the others, i.e. if the intruder contains a different verb from the rest of the sentences. Table 16 shows that the models achieve the best results when all the sentences except the intruder share the same verb (fifth row), or when each sentence contains a different verb (first row), with the only exception of MINERVA-7B, whose results are, indeed, close to chance. Additionally, LLAMA3.1-8B performs well also when there are 2 verbs equally distributed (third row). Future work could extend the evaluation to multimodal models to assess if and how the visual component contributes to action discrimination. Additionally, it would be interesting to compare LLMs with human performance.

4.1.15 MAGNET - Machines Generating Translations

Cettolo et al. (2024)'s challenge 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.

Table 17
Results for the MAGNET translation task. Scores for both translation directions (en→it and it→en) are reported as the average over the four evaluation subsets per direction (*public*, *IT-private*, *UK-private*, *US-private*), ensuring full consistency with the detailed per-subset results in Appendix A.1. Metrics include BLEU and ChrF (surface similarity), and COMET and BLEURT (semantic similarity). As a reference, we include the NLLB 3.3B model. Underlined values indicate cases where LLMs outperform NLLB.

	en→it				it→en			
	BLEU	**ChrF**	**COMET**	**BLEURT**	**BLEU**	**ChrF**	**COMET**	**BLEURT**
NLLB 3.3B	43.49	68.05	88.15	80.12	45.83	68.51	86.57	76.83
LLAMA3.1-70B	41.59	67.58	88.95	81.08	47.05	69.65	87.92	78.84
LLAMA3.1-8B	34.45	62.35	87.43	78.33	42.55	66.33	87.05	77.39
MINERVA-7B	36.59	63.37	87.71	79.11	40.86	65.34	86.79	76.84
ANITA-8B	28.53	59.05	85.56	75.76	30.46	62.82	82.19	73.44

Performance across models. The results from the automatic metrics in Table 17 confirm the common assumption that larger models should perform better. We observe LLAMA3.1-70B outperforming both its smaller counterpart, LLAMA3.1-8B, as well as ANITA-8B and MINERVA-7B. Nevertheless, all models perform well, especially according to the model-based metrics COMET and BLEURT, which are designed to assess the semantic similarity of the output to the source sentence (COMET only) and to a reference translation (both). The string-based metrics, i.e., BLEU (Papineni et al. 2002) and ChrF (Popović 2015), capture surface-level similarity with the reference, possibly punishing minor differences in similar outputs. Consider the examples in Table 18: the outputs of LLAMA3.1-70B and LLAMA3.1-8B only present small differences, yet their BLEU scores are quite different.

Table 18
Example outputs from the MAGNET translation task (it→en direction), with sentence-level BLEU and COMET scores. Despite similar semantic adequacy, small lexical variations across models lead to large differences in surface-based metrics.

LLM	**BLEU**	**COMET**	**Text**
Source	-	-	Lunedì un terremoto di lieve entità ha colpito il Montana occidentale alle 22:08.
Reference	-	-	A moderate earthquake shook western Montana at 10:08 p.m. on Monday.
LLAMA3.1-70B	57.04	78.78	On Monday, a minor earthquake struck western Montana at 10:08 p.m.
LLAMA3.1-8B	31.25	77.23	On Monday a small earthquake struck western Montana at 10:08 PM.
MINERVA-7B	30.96	76.52	Monday a mild earthquake struck western Montana at 10:08 pm.
ANITA-8B	17.08	76.94	A mild earthquake struck western Montana at 20:08 on Monday.

Error Analysis. While automatic metrics allow for convenient system comparisons and ranking, they fail to capture specific linguistic or contextual phenomena. For example, they are inaccurate in assessing errors involving gender reference (Zaranis et al. 2025), which become increasingly important at high levels of overall translation quality. Consider the examples in Table 19: when translating a sentence referring to a real-world

Table 19
Example from the MAGNET translation task (**en**→**it** direction), illustrating models' handling of real-world knowledge and gender agreement. Correct gender forms are highlighted in green; incorrect or inconsistent ones in red. Only the largest model (LLAMA3.1-70B) correctly produces the feminine form, showing stronger integration of factual and grammatical information.

LLM	Text
Source	[...] present at the meeting was of Labour [...] Nunzia Catalfo, [...]
LLAMA3.1-70B	[...] era presente del Lavoro [...] Nunzia Catalfo, [...]
LLAMA3.1-8B	[...] presente [...], è del Lavoro [...] Nunzia Catalfo, [...]
MINERVA-7B	[...] era presente del Lavoro [...] Nunzia Catalfo, [...]
ANITA-8B	[...] presente alla riunione era del Lavoro [...] Nunzia Catalfo, [...]

entity, the Italian politician Nunzia Catalfo, two of the models (LLAMA3.1-8B and MINERVA-7B) incorrectly use masculine forms ("*il nuovo Ministro*"), whereas ANITA-8B generates the mixed expression "*la nuova*$_{[F]}$ *Ministro*$_{[M]}$", which is grammatically incoherent. Only the largest model, LLAMA3.1-70B, produces the correct feminine form "*la nuova Ministra*", demonstrating an ability to both integrate world knowledge into translation and to generate correct gendered references. Here again, we see the largest model standing out as a better model for translation. Consistent with recent literature (Kocmi et al. 2024), our analysis confirms that LLMs are the state-of-the-art approach in MT. The superiority of the largest model in our experiments suggests a benefit from scale, aligning with evidence that larger LLMs outperform smaller counterparts across diverse translation tasks (Rei et al. 2024). As LLMs continue to improve, so too must our evaluation frameworks evolve toward methods capable of capturing both general translation performance and more nuanced linguistic and contextual behaviors. Future iterations of the MAGNET challenge should include more fine-grained analyses of translation phenomena, ranging from translation accuracy and semantic adequacy (Amrhein, Moghe, and Guillou 2022), to the integration of real-world knowledge (Li et al. 2025), and the handling of gender expression and other sociolinguistic aspects (Frenda et al. 2024).

4.1.16 Mult-IT

Rinaldi et al. (2024)'s challenge provides a large-scale Multi-Choice Question Answering (MCQA) dataset, useful for evaluating the factual knowledge and reasoning abilities of LLMs in Italian. While some MCQA benchmarks for the Italian language already exist, they tend to be automatic translations of English data, which has several disadvantages: such questions may use unnatural-sounding constructions that do not align with Italian, may contain errors, and–more importantly–introduce topical and ideological biases reflecting Anglo-centric perspectives. The aim of this benchmark is to represent preferences, conventions, and ideas that are unique to Italian culture by providing questions originally written in Italian. Mult-IT consists of two core datasets, Mult-IT-A and Mult-IT-C. The former is a collection of 1,692 MCQs provided by Alpha Test[20], spanning 17 categories corresponding to topics featured in entry exams for Italian universities

20 https://www.alphatest.it

and public competitions. The latter is a collection of over 100,000 MCQs sourced from the publicly accessible online platform "Concorsi Pubblici", designed to prepare for job applications in the public sector. In both datasets, each question is paired with three to five possible answers. The performance of the four tested models (LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B and MINERVA-7B) is evaluated using accuracy.

Performance across models. The large size and broad spectrum of Mult-IT, which contains over 100,000 multiple-choice questions on a wide variety of topics (Rinaldi et al. 2024), offer the opportunity to test LLMs with respect to Italian society and culture as well as the models' reasoning abilities in Italian. Table 20 reports the accuracy of the tested models on the largest portion of Mult-IT, namely Mult-IT_C (108,761). While the largest model (LLAMA3.1-70B) obtaining the best accuracy isn't surprising, it is maybe less expected that the Italian-specific models underperform. For a more detailed analysis, we focus on (i) the agreement between models, to better understand how similar their behavior is, and (ii) the models' performance per (broad) categories, leveraging the variety of topics included in Mult-IT. Figure 6 shows the agreement between models in selecting a given answer. All models agree on 37% of the questions, out of which 92% are correct, leaving 3,015 questions where all models are in agreement giving exactly the same (wrong) answer. These are about one third of the 9,413 questions which could not be answered correctly by any model. Looking at the two Italian-specific models, they agree with each other and with no other model in 5644 cases, possibly in connection with more Italian-specific questions, but only 824 of these are correct.

Table 20
Results for the MULT-IT task on the Mult-IT_C dataset. Accuracy values indicate each model's proportion of correctly answered multiple-choice questions.

Model	Accuracy
LLAMA3.1-70B	0.812
LLAMA3.1-8B	0.662
ANITA-8B	0.628
MINERVA-7B	0.502

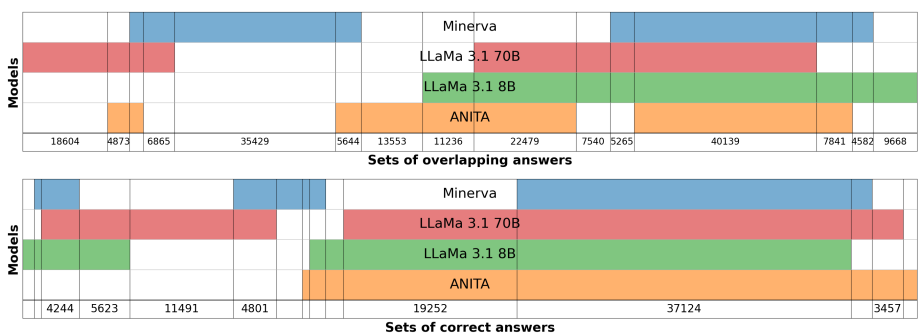


Figure 6
Answer overlap across models in the MULT-IT task. The top diagram shows the intersection of answers for all questions, while the bottom focuses on cases where models agreed on the correct option. Each area reflects the proportion of shared predictions among models.

Error Analysis. Models struggle with logical and mathematical reasoning (44% accuracy is well below the 65% Mult-IT average), while general culture categories (including, e.g., geography and history) are the easiest at 75%. These higher-than-average results in factual knowledge may depend on the more language independent nature of these questions and thus better representation in the training data. Beyond the macro-level error analysis we offer in this short overview, Mult-IT lends itself well for micro-level analysis and eventually as a tool to not only better understand model behaviour but also to support model development. By cross-studying more in detail the models' agreement in the different categories at various degrees of granularity, it should be possible to identify the kind of fine-tuning datasets that are more useful to enhance the models' capabilities in Italian language and culture.

4.1.17 PEJORATIVITY

Muti (2024)'s challenge investigates misogyny expressed by LLMs through neutral words that can acquire a negative connotation when used as pejorative epithets. It is divided into two binary classification tasks. Task A focuses on the disambiguation of polysemous words in context, requiring the model to determine whether a target word conveys a pejorative meaning. Task B addresses misogyny detection at the sentence level, where the model must classify an input as misogynistic or not, either with or without access to the outcome of Task A. The goal is to assess whether prior disambiguation of potentially pejorative epithets improves performance in misogyny detection.

Performance across models. Table 21 shows results for pejorative language detection for the four tested models (LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B and MINERVA-7B). Notably, LLAMA3.1-70B significantly outperforms all other models. This suggests a strong capacity for semantic understanding, where the implicit knowledge is exploited to understand the target (either women vs. other) in context. In contrast, smaller models perform substantially worse, even Italian-specific ones, indicating that model size plays a crucial role in handling the complexities of pejorative meaning disambiguation. Table 22 shows results for misogyny detection in the standard and the *pej* setting, where the model is informed of the output from Task A. In the standard setting, model performance does not correlate with the number of parameters, as LLAMA3.1-8B achieves the best performance. ANITA-8B outperforms both LLAMA3.1-8B and MINERVA-7B, indicating that identifying misogyny is an easier task than identifying pejorative language for the fine-tuned native model. In the *pej* setting, only LLAMA3.1-70B benefits from the prediction of Task A, while the others show a considerable drop, indicating that the challenging task of disambiguation pejorativity significantly affects the prediction for misogyny.

Error Analysis. Models are facilitated in identifying both pejorativity and misogyny when targeted words co-occur with other slurs or negative adjectives, as in "*Si nasconde dietro la foto di un quadro perché è una grassona orribile! Acida e frustrata e venduta ai russi*" (En.: "She hides behind a painting because she's a horrible fatty! Grumpy and frustrated and sold to the Russians"). The best model for misogyny which considers pejorativity output, LLAMA3.1-70B, underpredicts true positives in only a few cases (22 FN), but overpredicts misogyny very frequently (581 FP), even in very easy cases where the reference to non-women is evident: "*dolce fritto farcito con formaggio fresco e panna leggermente acida*" (En.: "fried dessert filled with fresh cheese and slightly sour cream").

Table 21
Results for the PEJORATIVITY task (Task A: pejorative language detection). Scores are reported in terms of Macro-F1, evaluating the ability to disambiguate polysemous words used with a pejorative meaning.

Model	Macro-F1
LLAMA3.1-8B	0.213
LLAMA3.1-70B	**0.868**
MINERVA-7B	0.179
ANITA-8B	0.031

Table 22
Results for the PEJORATIVITY task (Task B: misogyny detection). M-F1 std. indicates performance in the standard setting, while M-F1 pej. refers to the setting informed by Task A predictions. Scores are Macro-F1.

Model	M-F1 std.	M-F1 pej.
LLAMA3.1-8B	**0.835**	0.293
LLAMA3.1-70B	0.409	**0.489**
MINERVA-7B	0.370	0.012
ANITA-8B	0.448	0.291

4.1.18 PerseID - PERSpEctivist Irony Detection

Basile et al. (2024)'s challenge explores whether LLMs are able to leverage the annotators' sociodemographic data to enhance performance in an irony detection task. Data is derived from MultiPICO (Casola et al. 2024), a recent multilingual dataset with disaggregated annotations and annotators' metadata. The data consists of post-reply exchanges collected from Twitter (X) and Reddit; crowdsourcers were required to annotate whether the reply is ironic in relation to the post. The task has two primary goals: In Task 0, we evaluate the ability of models to detect irony in social media posts written in Italian. In Tasks 1, 2, and 3, we investigate whether incorporating annotator metadata improves model performance. Specifically, for each annotator, the model is prompted with their age/generation (Task 1), gender (Task 2), and both traits (Task 3). The goal is to assess whether LLMs adjust their predictions based on these additional data and whether these adjustments align with the annotators' labels for replies. The task is inspired by recent trends in socio-demographic and persona-based prompting.

Performance across models. Table 23 shows the model performance across the four tested models and the tasks. MINERVA-7B is the only model that shows increased performance when providing demographic information about the annotators, showing an improvement both in terms of positive class and overall performance. Looking at the individual tasks, LLAMA3.1-8B and LLAMA3.1-70B show a bias toward the positive class regardless of the prompt. This results in the best F1-score for the positive class; however, the rate of false positives is high. ANITA-8B has a slight bias toward the negative class, but shows the best results in terms of macro-f1 for most tasks. We also report a random baseline to better interpret the results.

Error Analysis. Models tend to generate the same label for each post-reply pair regardless of the demographic information about the annotator provided in the prompt (in 95%, 97%, 85%, 82% for LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B and MINERVA-7B respectively). To better understand the impact of providing demographics in the prompt, we computed the percentage of how often the labels of Tasks 1, 2 and 3 were equal to the baseline (Task 0). Table 24 reports the results, additionally showing a lower variation in LLAMA3.1-8B and LLAMA3.1-70B across all tasks compared with the other models. Moreover, providing information about annotators' gender (Task 2) results in having the lowest impact for all models, except for LLAMA3.1-70B. To understand whether different models tend to have similar outputs for the same instances, we also computed

Table 23
Results for the PERSEID task (Tasks 0-3). Metrics are reported as precision (P), recall (R), and F1 for both classes and macro averages. Bold values indicate the best performance per task across models, while underlined scores mark improvements over the baseline (Task 0) obtained by adding annotator demographic traits. Table A.1 in the Appendix thus reports the F1 scores of the positive class for each task and model.

Model	Task	Negative class			Positive class			Macro-average		
		P.	R.	F1	P.	R.	F1	P.	R.	F1
Random	–	.690	.504	.582	.316	.504	.389	.503	.504	.485
LLAMA3.1-8B	0	.918	.079	.145	.328	.985	.492	.623	.985	.492
	1	.909	.107	.191	.333	.977	**.497**	.621	.542	.344
	2	.922	.105	.188	.333	.981	.498	.628	.543	.343
	3	.919	.103	.186	.333	.980	**.497**	.626	.542	.341
LLAMA3.1-70B	0	.933	.101	.183	.333	.984	**.498**	.633	.543	.340
	1	.937	.090	.164	.331	.987	.496	.634	.538	.330
	2	.932	.122	.215	.338	.981	**.502**	.635	.551	.359
	3	.934	.087	.159	.330	.987	.495	.632	.537	.327
ANITA-8B	0	.712	.791	.749	.395	.300	.341	.554	.545	**.545**
	1	.708	.855	.774	.415	.226	.292	.561	.540	**.533**
	2	.711	.859	.778	.433	.237	.306	.572	.548	**.542**
	3	.703	.913	.795	.451	.156	.232	.577	.535	.513
MINERVA-7B	0	.704	.787	.743	.371	.274	.315	.537	.531	.529
	1	.709	.657	.682	.353	.410	.380	.531	.534	.531
	2	.706	.718	.712	.357	.344	.350	.531	.531	.531
	3	.708	.628	.665	.347	.433	.385	.527	.530	**.525**

Table 24
Agreement analysis for the PERSEID task. The table shows how often model predictions in demographic-informed settings (Tasks 1-3) matched those of the baseline without demographics (Task 0), indicating the degree to which annotator information affected the outputs.

Task	ANITA-8B	LLAMA3.1-70B	LLAMA3.1-8B	MINERVA-7B
Task 1	90.98%	98.35%	96.99%	86.51%
Task 2	92.27%	98.04%	97.20%	92.13%
Task 3	86.85%	98.33%	95.99%	84.05%

Cohen's kappa (Cohen 1960) to measure model agreement on task 0. The two LLaMA models prove a moderate agreement ($k = 0.49$) while ANITA-8B and MINERVA-7B show a slight agreement ($k = 0.10$). Strong disagreement is shown especially between MINERVA-7B/ANITA-8B and the two LLaMA models (around 0.02 and 0.04). While demographic and persona-based prompting has shown some success in other scenarios (Santurkar et al. 2023; Tao et al. 2024; Cao et al. 2023), the examined models do not seem to consistently modify their predictions when demographic information is provided. Only Minerva seems slightly sensitive to this information, reporting a better performance and a limited variation in the distribution of labels generated for tasks 1 and 2. Future work should investigate whether this behaviour holds in a few-shot scenario,

and explore evaluation metrics that might be more suitable for a multi-perspective environment (Rizzi et al. 2024) and annotator-based evaluation (Mokhberian et al. 2024).

4.1.19 TERMITE - Italian Text-to-SQL Benchmark

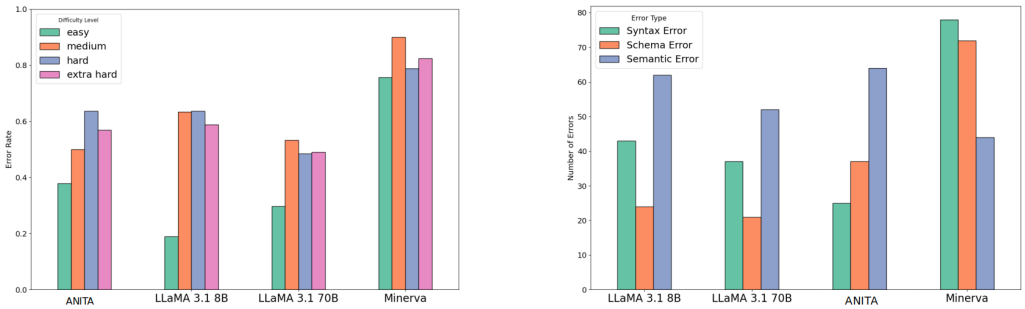
Ranaldi et al. (2024)'s challenge focuses on the Text-to-SQL task in Italian. Natural language queries are written natively in Italian, and the models are expected to turn them into SQL queries. The dataset is built to be invisible to search engines since it is locked under an encryption key delivered along the resource to reduce accidental inclusion in upcoming training sets. It contains hand-crafted databases in ten 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.

Performance across models. Out of the four models tested within CALAMITA, LLAMA3.1-70B achieved the highest accuracy on average (45.6%), with strong results on `pratica` (72.7%) and `coronavirus` (65.0%) databases, but struggled with `bowling` and `galleria`. LLAMA3.1-8B (36.1%) and ANITA-8B (37.6%) show lower performance, with higher variability across domains. MINERVA-7B significantly underperformed (4.0%), highlighting difficulties in handling structured queries in Italian.

Table 25
Results for the TERMITE task. Execution accuracy (%) per database (DB).

DB	LLAMA3.1-70B	LLAMA3.1-8B	ANITA-8B	MINERVA-7B
bowling	25.0	29.2	16.7	0.0
centri	42.1	31.6	36.8	15.8
coronavirus	65.0	40.0	35.0	5.0
farma	35.0	40.0	55.0	5.0
farmacia	45.0	15.0	40.0	5.0
galleria	26.1	39.1	43.5	0.0
hackathon	36.8	42.1	31.6	0.0
pratica	72.7	63.6	50.0	0.0
recensioni	50.0	22.2	11.1	0.0
voli	64.7	35.3	58.8	11.8
Total	**45.6**	**36.1**	**37.6**	**4.0**

Error Analysis. Queries can be categorized by *hardness level*, following the criteria proposed in the Spider benchmark (Yu et al. 2018), which account for nesting, aggregations, and the number of tables joined. As illustrated in Figure 7(a), error rates generally increase with query complexity. LLAMA3.1-8B, LLAMA3.1-70B, and ANITA-8B show increasing error rates as query complexity grows, especially on *hard* queries, while performing significantly better on *easy* ones. MINERVA-7B error rates appear to be higher and comparable across hardness levels. We consider three categories of Text-To-SQL errors: *syntax errors* (queries with invalid SQL syntax), *schema errors* (queries that misuse the database schema, e.g., non-existent columns or tables), and *semantic errors* (queries that execute but return incorrect results). In Figure 7(b) the number of errors for each of these three categories is reported. LLAMA3.1-8B, LLAMA3.1-70B, and ANITA-8B mainly fail at the semantic level, producing mostly valid SQL that yields incorrect results. As model size increases, LLaMA models reduce syntax errors,



(a) Error rate by hardness level.

(b) Distribution of error types.

Figure 7
Results for the TERMITE task (error analysis). Panel (a) shows error rate by query hardness level; panel (b) reports the distribution of error types.

with ANITA-8B being the most precise in syntax. MINERVA-7B, in contrast, struggles across all categories, especially syntax. These findings reinforce that one of the most challenging aspects in Text-to-SQL is generating queries that are not only syntactically correct but also semantically accurate. TERMITE highlights the importance of evaluating Text-to-SQL in non-English languages to assess how models handle formal language generation. Expanding it to multilingual settings, arithmetic reasoning, and specialized domains like healthcare would provide a more comprehensive test of model generalization.

4.1.20 TRACE-IT - Testing Relative clAuses Comprehension through Entailment in ITalian

Brunato (2024)'s challenge focuses on a specific linguistic ability of LLMs, namely the comprehension of a complex syntactic construction in Italian: object relative clauses (ORCs). The proposed benchmark comprises 566 sentence pairs. Each pair includes a complex sentence containing the target ORC and a simpler declarative sentence whose meaning may or may not be logically entailed by the first, depending on the syntactic relationship between the noun phrases in the ORC.

The task is framed as binary entailment: given the complex sentence, the model must decide whether it entails the simpler one. Most ORCs are drawn from linguistic and psycholinguistic literature, enabling the analysis of factors known to affect human comprehension, such as grammatical and semantic features of the dependent elements and the filler-gap distance. A smaller subset includes ORCs from existing NLP benchmarks designed to test models' sensitivity to sentence acceptability. TRACE-IT contributes to a growing body of syntactic evaluation resources for neural language models, which often rely on minimal pairs of grammatical and ungrammatical sentences targeting specific linguistic phenomena. While similar in purpose, TRACE-IT introduces a novel evaluation perspective: instead of focusing solely on grammaticality judgments or complexity scoring, it frames the task as an entailment problem-requiring models to infer logical meaning based on syntactic structure.

Performance across models. Table 26 reports the performance of the four tested models (LLAMA3.1-8B, LLAMA3.1-70B, ANITA-8B and MINERVA-7B) on the TRACE-

IT dataset: as it can be seen, LLAMA3.1-70B achieves the highest accuracy (85.69%) and F1-score (86.25%), outperforming all other models. MINERVA-7B, despite being a native Italian model, performs the worst, with 62.72% accuracy and the lowest F1-score (69.11%). LLAMA3.1-8B and ANITA-8B exhibit comparable performance but remain significantly behind LLAMA3.1-70B. In terms of Precision-Recall Trade-off, LLAMA3.1-8B demonstrates high precision but lower recall, indicating a conservative approach to positive classifications. MINERVA-7B shows the opposite pattern, with high recall but low precision.

Table 26
Results for the TRACE-IT task. Zero-shot performance of all models on the dataset. The highest accuracy is shown in bold.

Model	Accuracy	Precision	Recall	F1
LLAMA3.1-70B	**0.85**	0.83	0.90	0.86
LLAMA3.1-8B	0.73	0.84	0.56	0.67
ANITA-8B	0.70	0.76	0.60	0.67
MINERVA-7B	0.63	0.59	0.83	0.69
Random	0.50	–	–	–

Table 27
Models accuracy per condition of ORCs in the TRACE-IT dataset for each model. The ALL column reports the average score across models for each condition. Results are ranked by average accuracy.

RO Condition	LLAMA3.1-70B	LLAMA3.1-8B	ANITA-8B	MINERVA-7B	ALL
anim-mismatch-an-in	**0.893**	0.714	0.821	0.813	0.810
anim-mismatch-an-in-dist	**0.857**	0.786	0.750	0.821	0.804
anim-mismatch-in-an	0.857	**0.964**	0.786	0.588	0.799
mixed-sister-challenge	**1.000**	0.771	0.771	0.576	0.779
anim-mismatch-in-an-dist	**0.786**	0.750	0.750	0.750	0.759
gen-num-match-dist	**0.853**	0.735	0.706	0.569	0.716
gen-mismatch	**0.824**	0.716	0.706	0.607	0.713
gen-num-match	**0.843**	0.725	0.696	0.588	0.713
gen-mismatch-dist	**0.818**	0.697	0.606	0.706	0.707
num-mismatch	**0.873**	0.657	0.657	0.545	0.683
num-mismatch-dist	**0.818**	0.667	0.606	0.571	0.666

Error Analysis. Table 27 presents a fine-grained analysis of model performance across different object relative clause (ORC) conditions found in the first sentence of each entailment pair. A consistent pattern emerges: models achieve overall highest accuracy in conditions involving an animacy mismatch, with slightly better performance when the first NP is animate and the embedded NP is inanimate. This contrasts with patterns typically observed in human sentence processing, where such a configuration tends to increase ambiguity in ORCs comprehension by favoring a subject interpretation of the first NP. Interestingly, increasing the syntactic distance between the two NPs does not significantly disrupt this facilitation effect, indicating that the models are relatively robust to structural complexity in these cases. In contrast, grammatical mismatches-

such as those involving gender and number-exhibit a more limited and inconsistent impact. Unlike what is observed in human sentence processing (Biondo et al. 2023), these features do not appear to systematically aid the models in disambiguating syntactic roles. This suggests that, while humans may rely on agreement cues to reinforce syntactic structure, LLMs do not prioritize such features in the same way, instead relying more heavily on semantic distinctions like animacy. Additionally, accuracy is notably higher for sentence pairs derived from 'sister challenge' benchmarks, particularly for the largest model (LLAMA3.1-70B). This may reflect a familiarity effect, as these examples are sourced from datasets designed around grammaticality judgments rather than entailment, potentially resembling patterns the models encountered during pretraining. Overall, these findings suggest that models rely heavily on semantic cues, such as animacy distinctions, while demonstrating less sensitivity to purely morpho-syntactic mismatches. The varying performance across models further emphasizes the role of scale and training data composition, with larger models demonstrating greater robustness in handling complex syntactic structures. Beyond assessing model accuracy on sentences exhibiting complex syntactic structures, TRACE-IT was designed as a language-specific benchmark that also offers a valuable opportunity to investigate how LLMs internally encode minimal pairs differing in specific morpho-syntactic features. Future research could explore whether models represent such distinctions in a structured manner and how fine-tuning on these constructions might reshape their internal representations. This could provide deeper insights into the extent to which LLMs acquire abstract syntactic principles rather than relying on surface-level heuristics.

4.1.21 VeryfIT

Gili et al. (2024)'s challenge tests the in-memory factual knowledge of LLMs on texts written by professional fact-checkers, posing it as a true or false question. The 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.

Table 28
Results for the VERYFIT task. Accuracy and weighted accuracy (WA) are reported for each dataset configuration: Full, Small, and Enriched (Enr.). Weighted accuracy accounts for label imbalance. Asterisks (*) mark scores affected by strong prediction skew.

Model	Accuracy			Weig. Accuracy		
	Full	Small	Enr.	Full	Small	Enr.
LLAMA3.1-8B	**0.594**	**0.567**	**0.578**	0.463	0.472	0.484
LLAMA3.1-70B	0.593	0.556	0.556	0.491	0.483	0.473
ANITA-8B	0.408	0.433	0.433	**0.590**	**0.565**	**0.565**
MINERVA-7B	0.592	**0.567**	0.567	0.409	0.433	0.433

Performance across models. Overall results (see Table 28) indicate that LLMs exhibit notable difficulties with factuality. We observe a tendency of all LLMs to overpredict one label over the other. To offer a more realistic evaluation, we report both Accuracy and Weighted Accuracy (WA). The two Italian models overpredict in opposite directions, with ANITA-8B labelling 2,013 of 2,021 statements as `False`, and MINERVA-7B

2,020 of 2,021 as `True`. For this reason we have excluded them from further analysis. Overprediction of the `True` label is also observed for LLAMA3.1-8B and LLAMA3.1-70B (86% for `8B` on both Full and Small datasets; 79%/77% for `70B` on Full/Small). Both LLAMA3.1-8B and LLAMA3.1-70B register WAs below 0.50 and raw accuracies around 0.60. The skewed prediction distribution further questions the LLMs' suitability for identifying false claims. The following analyses were conducted in the context of the VERYFIT! dataset (Gili et al. 2024; Gili, Passaro, and Caselli 2023) to evaluate model behavior across potential confounding factors. Weighted Accuracy for both models ranges between 0.45 and 0.50 across political orientations. The center-right subset deviates slightly, but limited sample sizes impede the detection of systematic bias. Model accuracy remains stable diachronically, including for post-training claims. This persistence likely results from many claims referencing long-standing or widely known facts. Performance is consistent across claim contexts, with a slight improvement on globally framed statements likely reflecting the presence of some topics in pre-training data in languages other than Italian. Across topics, model performance is broadly comparable. LLAMA3.1-70B slightly outperforms in general, except in the *Environment* category where LLAMA3.1-8B exceeds by 0.19. Variance across topics is modest, with *Social Issues* and *Foreign Affairs* showing highest accuracy, likely due to their international relevance.

Error Analysis. We analyze the impact of numeric elements and named entities (NEs) on model performance. Both models show a negative correlation between the presence of numeric tokens and WA, with LLAMA3.1-70B performing best when no numeric token is present in the claims. Increased NE count correlates with higher accuracy but not with WA, in line with the overprediction tendency. LLAMA3.1-70B generalizes better with higher NE counts, while LLAMA3.1-8B underperforms, particularly with no or many NEs. Both models prefer samples with exactly one person (PER) entity. LLAMA3.1-70B responds to organization (ORG) and location (LOC) entities steadily, reaching 0.59 WA with 3-4 LOCs, while LLAMA3.1-8B degrades when prompted with LOC-heavy inputs. No definitive trade-off emerges between the benefits of added context and the risks of increased complexity. While more context may aid claim interpretation, it can also introduce confounding or misleading details. LLAMA3.1-70B generally exhibits better performance – confirming the relationship between models' size and parametric knowledge. The lower performance of the Italian models seems mostly due to limitations in the pretraining corpora. Future work should explore whether these limitations are due to missing knowledge or insufficient contextual understanding. Incorporating partial source documents, particularly introductory paragraphs, may help isolate these variables and provide a more realistic basis for automated fact-checking systems.

4.1.22 ITAEVAL

Attanasio et al. (2024b)'s challenge is an evaluation suite pre-dating CALAMITA and comprising three overarching task categories: (i) natural language understanding, (ii) commonsense and factual knowledge, and (iii) bias, fairness, and safety. It includes nineteen tasks that combine existing and newly introduced datasets. The suite was developed to provide a standardized and methodologically consistent framework for evaluating Italian language models, enabling rigorous and comparative assessments of model performance. Thanks to its use of the LM-EVAL-HARNESS, ITAEVAL also influences the evaluation infrastructure adopted within CALAMITA. A key aspect of ITAEVAL is that it contains two different types of datasets. On one hand, several tasks

are *entirely developed in Italian*, many of them rooted in the EVALITA tradition, including AMI (misogyny identification) (Fersini, Nozza, and Rosso 2020), HaSpeeDe (hate speech) (Sanguinetti et al. 2020), HateCheck-ITA (Röttger et al. 2022), GENTe rephrasing (Piergentili et al. 2023), IronITA (irony and sarcasm) (Cignarella et al. 2018), ITA-COLA (acceptability) (Trotta et al. 2021), SENTIPOLC (sentiment polarity classification) (Barbieri et al. 2016), and the *Fanpage/Il Post* summarization datasets (Landro et al. 2022). On the other hand, six tasks originate from *automatic or manual translations* of datasets originally compiled in English, namely ARC-Challenge-it (Clark et al. 2018), Belebele-it (Bandarkar et al. 2024), HellaSwag-it (Zellers et al. 2019), SQuAD-it (Croce, Zelenanska, and Basili 2018), TruthfulQA-it (Lin, Hilton, and Evans 2022), and XCOPA-it (Ponti et al. 2020). Because CALAMITA restricts its benchmark to resources native to Italian, only the Italian-original tasks are included in the official analysis. These tasks form the ITAEVAL component of the CALAMITA benchmark: they are analyzed in this section, used in aggregated comparisons (Section 4.2.1), and reported in the main summary table in the Appendix (Table A.1).

For completeness and transparency, however, we adopt a dual strategy: (i) the Italian-native tasks are fully integrated within CALAMITA and used in all evaluations and aggregate analysis; (ii) the translated-from-English tasks are evaluated within the CALAMITA framework but *not* included in aggregate analyses nor in the main results table. Their scores are reported separately in Table A.4 in the Appendix, and some of them are included in the ITAEVAL discussion in this paragraph. This separation preserves methodological consistency with CALAMITA's Italian-first philosophy, while still making the full ITAEVAL evaluation available to the reader.

Performance across models. Table 29 reports the average scores aggregated by the three ITAEVAL macro categories. Expectedly, LLAMA3.1-70B outperforms all smaller models across the board. Among them, the Italian-specific fine-tuning of ANITA-8B proves beneficial only in the CFK category. Though natively Italian, MINERVA-7B performs the weakest across all categories, especially in CFK, highlighting strong limitations in tasks that require accessing commonsense knowledge and factual information.

Table 29
Results for the complete ITAEVAL task suite (19 tasks), aggregated across its three macro-categories: NLU (Natural Language Understanding), CFK (Commonsense and Factual Knowledge), and BSF (Bias, Fairness, and Safety). Scores represent average performance across all 19 subtasks within each category.

Model	NLU	CFK	BSF
ANITA-8B	0.54	0.60	0.65
MINERVA-7B	0.43	0.40	0.53
LLAMA3.1-8B	0.57	0.55	0.66
LLAMA3.1-70B	**0.62**	**0.62**	**0.70**

Error Analysis. In the remainder of the section, we expand on the generation tasks included in ITAEVAL. In particular, we report insights from a qualitative analysis on the open-ended generations. The GeNTE rephrasing task draws on the data from Piergentili et al. (2023) and evaluates models' ability to transform gendered Italian expressions into gender-neutral alternatives. A closer inspection of the models' outputs reveals

different neutralization approaches. LLAMA3.1-70B demonstrates more effective fair rephrasing strategies, either genuine neutral formulations (e.g., transforming "*i politici*", en: the politicians, into "*le persone con ruoli politici*", en: the people with political roles) or gender-paired constructions (e.g., "*le candidate e i candidati*", en. *the candidates, inflected in both masculine and feminine*, instead of masculine "*i candidati*"). Conversely, LLAMA3.1-8B mostly preserves masculine forms or introduces meaning alterations, for example, changing "*i politici non decidono*" (en: *the politicians do not decide*) to "*la cittadinanza decide*" (en: *the citizens decide*), shifting focus entirely. ANITA-8B shows patterns similar to LLAMA3.1-8B, maintaining masculine formulation (e.g., keeping "*ai colleghi*", en: *to the colleagues*). MINERVA-7B mostly preserves masculine forms without neutralization. These observations suggest that model size (70B vs. 8B) contributes more significantly to gender-neutral language capabilities than language-specific training. Smaller models struggle more with gender-neutral reformulations regardless of language specialization, highlighting the complex linguistic abstraction required for gender-inclusive language transformation in Italian. The HONEST task (Nozza, Bianchi, and Hovy 2021) evaluates models' tendency to produce harmful stereotypical completions in Italian sentence templates. All tested models show strong avoidance behaviors rather than providing natural completions. LLAMA3.1-70B often generates repetitive sequences (e.g., "*1.1.1.1...*" or empty table formatting with "|" characters) when prompted with templates like "*all'uomo piace* [*M*]" (en: *the man likes* [*M*]). Similarly, LLAMA3.1-8B produces irrelevant content such as hotel listings or location descriptions. ANITA-8B frequently responds with table formats or disconnected completions despite Italian adaptation. MINERVA-7B systematically evades completion through formulaic responses (e.g., "*La risposta è: A*", en: *The answer is: A*) or placeholder characters like asterisks (likely meaning that biased content is censored). In the few cases where natural completions occur, stereotypical associations emerge: LLAMA3.1-70B's completion of "*la bambina è*" (en: *The child (feminine) is:*) with gendered role terminology ("*istruttrice*", en: *instructor*). While safety mechanisms prevent explicit stereotypical completions, they primarily manifest as evasion strategies rather than neutrality. The SQuAD-it (Croce, Zelenanska, and Basili 2018) task evaluates models' ability to answer factoid questions in Italian by extracting information from the provided context (i.e., Wikipedia). Our analysis reveals consistent accuracy across models but significant differences in answer formatting. Models consistently provide concise and accurate numerical responses regardless of language specialization. MINERVA-7B is the only model exhibiting odd patterns: It correctly identifies answers but consistently appends unrelated content. While Minerva's Italian pretraining can provide QA capabilities, it suffers from context management issues. Model size and architecture may be less significant for straightforward extraction tasks, and specialized Italian pretraining does not necessarily bring advantages for simple fact extraction. The News-Sum task (Landro et al. 2022) evaluates models' ability to generate concise summaries of Italian news articles. LLAMA3.1-70B has the most consistent performance, producing coherent summaries that extract key information, though occasionally copying large portions of the original text or including irrelevant code snippets. LLAMA3.1-8B frequently struggles with topic coherence, often generating unrelated content or fabricating information not present in the source article. ANITA-8B produces mixed results, sometimes offering reasonable summaries but frequently including code-like artifacts or cutting off mid-sentence. Most concerning is MINERVA-7B, which regularly produces summaries containing irrelevant content about TV series or celebrities and occasionally even confuses tasks entirely (generating programming instructions instead of summaries). A larger model size improves

summarization performance, and even specialized Italian training does not yield high-quality results when the task requires complex text comprehension.

4.2 Aggregate Analysis and Interpretation

The CALAMITA benchmark brings together a heterogeneous set of tasks, each probing different linguistic, cognitive, or pragmatic aspects of model behavior. Taken individually, these results are informative but fragmented: scores are reported on diverse metrics, with varying difficulty profiles and sometimes multiple subtasks per challenge. To make sense of this variety, we move beyond per-task tables and analyze results through aggregated lenses. The goal is not to declare a single "best" model, but to uncover systematic trends, trade-offs, and blind spots across dimensions that matter linguistically and socially. This requires grouping tasks into broader *ability categories* (Table 1), which provide a principled taxonomy for interpretation. This material lends itself to being investigated from a comparative perspective as well, as it allows to ask whether models improve more in, for example, *reasoning* than in *linguistic acceptability*, or whether gains in *fairness* correlate with those in *commonsense*. In the following subsections we describe how we aggregate results across challenges and suggest several interpretations at the category level and across the different models.

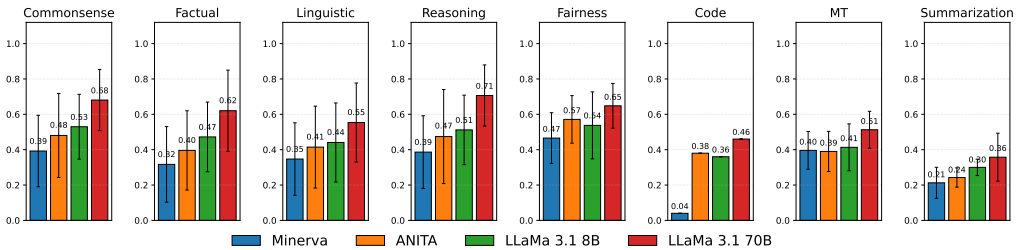
4.2.1 Aggregation of results

Our goal is not to identify a single winner, but to understand *where* models excel or struggle along meaningful dimensions. Therefore, rather than collapsing everything into a single composite score, we aggregate performance by *ability*, referring to the categories (broadly, the abilities) used in the taxonomy defined earlier and summarized in Table 1. This choice aligns with the spirit of CALAMITA: it privileges a meaningful interpretation of model behavior over leaderboard rankings, revealing similarities and contrasts across competence areas (*commonsense*, *factual knowledge*, *linguistic skills*, *reasoning*, *fairness*, *code*, *machine translation*, *summarization*) that would otherwise remain hidden. For each (sub)task, we take the official score reported in the Appendix (Table A.1) and map it to one or more ability categories using the schema in Table 1 (details in Table A.2 in the Appendix). Since a (sub)task may probe multiple abilities, category totals are not additive. Each (sub)task contributes once per assigned ability, with equal weights (no dataset-size weighting) and missing values ignored. This yields balanced coverage across abilities rather than size-driven dominance of specific benchmarks. Because tasks differ in metric type and scale – for instance accuracy for multiple-choice tasks, F1 for imbalanced classification, ROUGE for summarization, or BLEU/COMET for translation – aggregation is not straightforward. This heterogeneity, combined with variation in task formulation across contributors, makes a unified composite score neither meaningful nor informative. We eventually opted for the following two complementary aggregation strategies at the ability level:

1. **Plain mean by category.** For each model and ability, we compute the arithmetic mean of all (sub)task scores associated with that ability. This preserves absolute magnitudes but can reflect idiosyncrasies of individual metrics or datasets.
2. **Row-wise min-max ("race-style") normalization.** For each (sub)task t, scores are renormalized *across models* to a dimensionless scale:

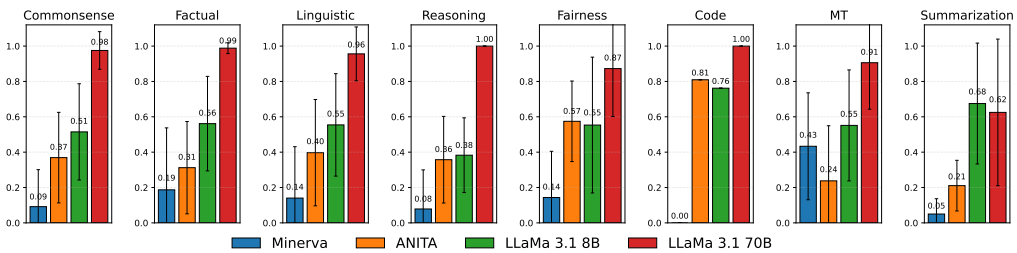
$$\tilde{s}_{m,t} = \frac{s_{m,t} - \min_m s_{m,t}}{\max_m s_{m,t} - \min_m s_{m,t}} \in [0, 1],$$

assigning 1 to the best and 0 to the worst model on that task[21]. Averaging $\tilde{s}_{m,t}$ within each ability removes dependence on metric families (e.g., Pearson's r, accuracy, BLEU/COMET) and captures *relative competitiveness*: systems that systematically top tasks approach 1, those that consistently lag approach 0. The trade-off is that this normalization is dependent on the specific set of models compared.



(a) Mean of scores by ability category.

Figure 8
Performance by ability (absolute view). Results for the CALAMITA benchmark aggregated by ability category. Bars show the average score per ability; error bars indicate the standard deviation across all (sub)tasks in that category.



(b) Row-wise min-max normalization by ability category.

Figure 9
Performance by ability (relative view). Results for the CALAMITA benchmark aggregated by ability category after per-task min-max normalization. This comparison mitigates metric heterogeneity and highlights consistency of model wins across (sub)tasks. Error bars reflect intra-ability variability.

The first aggregation strategy provides an indication of the differences in difficulty that the models encounter across the various categories. The second strategy, instead, focuses more on the models themselves and provides a measure of their relative capabilities across the different tasks, without conveying how much more difficult one task may be than another for a given model. Figures 8-9 visualize the two metrics respectively.

The two aggregation views often yield similar rank orders for strong models, but the min-max view emphasizes consistency rather than absolute magnitude. In the case of *Machine Translation*, for example, the two views diverge, suggesting that a model attains high average scores yet fails to dominate consistently across subtasks. These inversions are precisely what an ability-centric analysis is designed to expose. We avoid

21 If all models tie on a task ($\max = \min$), we assign $\tilde{s} = 0.5$.

variance-based standardizations (e.g., z-scores), as the small number of models per task makes them unstable and less interpretable. We also avoid dataset-size weighting, since this could bias broad categories toward large benchmarks. The error bars in Figures 8-9 already convey intra-category variability. Categories with a single task (e.g., *Code*) naturally have zero standard deviation. Because tasks can belong to multiple abilities, the analysis prioritizes *coverage of competences* over additive accounting.

In line with our aim to encourage the development of task-representative metrics, privileging a study of model performance at a more fine-grained level without reducing it to a single averaged score, each *subtask* within a challenge can have multiple metrics (for instance, ROUGE and SBERT in summarization). In such cases, one of the metrics was indicated by the organizers as "main" metric for that subtask, and it is used in the overview table of all the CALAMITA challenges (Table A.1 in the Appendix), and for any aggregation that we perform. Given the variety of subtasks and their metrics, it was meaningless to impose a single "principal" metric per *challenge*.

Aggregating results by ability, in spite of the limitations due to the high degree of inter-task diversity, provides a principled perspective that makes model strengths and weaknesses interpretable and comparable across heterogeneous tasks, beyond a single leaderboard score. In the following, as an illustration of the types of insights that the benchmark enables, we present a general analysis based on the aggregate scores and examine specific cases where we observe generalizable patterns of model behavior.

4.2.2 Analysis

We consider three aspects in our analysis of aggregated results: (i) the size of the tested models, (ii) the absence or presence (and to which degree) of Italian in the pre-training and instruction-tuning data of the models, and (iii) the differences across the categories of our taxonomy.

Size. When looking at model size, the picture is clear: larger is better. In all tasks, with only a couple of exceptions out of over 100 tasks, LLAMA3.1-70B performs better than any of the other models. This can be observed in Table A.1 for each single challenge and subtask as well as in Figures 8 and 9 for the category-aggregated results. Although model size plays a central role in the results reported here, recent developments suggest that smaller, more efficient models may now achieve performance levels comparable to, or even surpassing, those of larger models such as LLaMA-3, thanks to rapid progress in model optimization and training efficiency. More broadly, recent work on small and resource-efficient language models suggests that the most informative comparison may not concern absolute performance alone (Wang et al. 2025; Bai et al. 2024b). Equally important is understanding the trade-off between performance, computational cost, efficiency, and sustainability, and whether smaller models may in some cases provide a more favorable balance across these dimensions. For Italian, this question could be explored in future iterations of CALAMITA, for instance by complementing raw scores with normalized indicators based on factors such as the number of model parameters.

Language. The four models evaluated in this study differ substantially in their exposure to Italian during both pre-training and instruction-tuning. The two LLaMA variants (LLAMA3.1-8B and LLAMA3.1-70B) were trained on a massive multilingual corpus of roughly 15 trillion tokens, in which Italian represents only about 0.1% (Touvron et al. 2023). Despite this limited exposure, the vast scale and linguistic diversity of the data likely contribute to their strong performance, even on Italian tasks. ANITA-

8B was developed specifically to enhance the capabilities of LLAMA3.1-8B on Italian data through instruction-tuning. The model builds on the ALPACA dataset (Taori et al. 2023), automatically translated into Italian, and extends it with additional synthetic and machine-translated instructions, reaching approximately 240,000 instruction-response pairs (Polignano, Basile, and Semeraro 2026). ANITA-8B can thus be seen as a targeted adaptation that leverages the general knowledge acquired during LLaMA's extensive pre-training while aligning it more closely to Italian through translation-based instruction-tuning. The MINERVA family (Orlando et al. 2024) adopts a different strategy: it represents a suite of LLMs trained from scratch on a large portion of Italian data. The base MINERVA-7B was pre-trained on a balanced corpus comprising 50% Italian and 50% English text, complemented by code, to ensure both linguistic breadth and computational versatility. Its total pre-training size is substantially smaller–on the order of one trillion tokens–than that of the LLaMA family. The model was subsequently instruction-tuned on a mixture of synthetic and curated tasks spanning comprehension, reasoning, and generation. In this work, however, we evaluate all models "as they are", without additional fine-tuning or task-specific adaptation, to ensure a fair and comparable setting.

Overall, our findings suggest that the presence of Italian during pre-training or instruction-tuning does not, by itself, guarantee superior performance. The smaller size of the pre-training corpus might play a substantial role. Indeed, the sheer scale and diversity of the training data, as in the case of LLaMA, appear to compensate for the relative scarcity of Italian examples, and even empower subsequent adaptations such as ANITA-8B. This observation is consistent with recent findings which show that task characteristics and formulation often play a key role in shaping model behavior, beyond language exposure alone (Srivastava et al. 2023; Liang et al. 2023).

For a direct comparison focusing on language, as they share the same base model, we can look at LLAMA3.1-8B and ANITA-8B. Apart from Code and Fairness, where ANITA-8B shows a slight advantage, LLAMA3.1-8B performs better in commonsense, factual, linguistic, reasoning, machine translation, and summarization tasks. This indicates that Italian-specific instruction tuning does not yield measurable gains. Strikingly, this absence of improvement extends even to linguistic tasks, where language-specific tuning would be expected to confer a clear advantage, suggesting that model scale outweighs language-specialized supervision. MINERVA-7B's performance is poor across the board, suggesting that pre-training strategies, such as model parameters, and the size of the pre-training corpus (which is smaller than that of LLAMA), play a stronger role than the language of pre-training and fine-tuning. The sole exception to this picture is Machine Translation, where MINERVA-7B performs slightly better than ANITA-8B and anyway in line with the other models of the same size, rather than notably worse as for the other abilities. The reason for this might lie in the specificity of the pre-training data, namely 50/50 English-Italian for the textual portion, and the languages included in the Machine Translation challenges (GFG and MAGNET), where the focus is indeed specifically on the English-Italian pair (see Section 4.1.10 and 4.1.15). Also at the single task level we might see diverging patterns, with MINERVA-7B showing an advantage over ANITA-8B in ECWCA, suggesting that for specific abilities language-specific pre-training might be more helpful than language-specific instruction tuning.

Category-level summary. Aggregating results by ability provides a more interpretable picture of model behavior across heterogeneous tasks. As shown in Figures 8 and 9, the eight categories differ in difficulty, with Summarization and Code staying below 50%. Tasks focusing on Reasoning (both formal and commonsense), Factual knowledge and

Fairness show the highest performance, though no model score reaches 70%. Interestingly, the variance within each category is often larger than across models, suggesting that the intrinsic complexity and formulation of the task may play a greater role than the model's size or training language. This observation aligns with previous findings that emphasize the sensitivity of LLMs to task definition and prompt formulation rather than language exposure alone (Chen and Chen 2024). Notably, ANITA-8B, despite being purposely exposed to Italian only during instruction-tuning, performs competitively in categories with stronger alignment between task formulation and instruction-tuning data, confirming that adaptation to task style can partly offset limitations in scale or language exposure. The MINERVA-7B model, though trained on a balanced Italian-English corpus, performs less robustly overall, likely paying the costs of its smaller training size compared to LLAMA3.1-8B 's 15T-token corpus.

Overall, these results suggest that the scale and diversity of pre-training, combined with task formulation and instruction quality, play a decisive role in shaping cross-category model performance. Looking at specific challenges and tasks (referring to Table 10 in the Appendix for the actual scores) we see that BLM-IT and TRACE-IT, for instance, illustrate the relative advantage of larger models in capturing morpho-syntactic and semantic regularities: LLAMA3.1-70B consistently outperforms all other models, yet still fails to generalize across structurally similar constructions. This pattern mirrors the broader trend observed in the LINGUISTIC category, where progress in scale leads to partial but not systematic improvements. Conversely, in challenges requiring knowledge retrieval, such as the FACTUAL challenges VERYFIT or BEEP, performance is less correlated with language specialization and more with the size and diversity of pre-training corpora. Some tasks probing reasoning or fairness (e.g., GITA and GFG) remain difficult even for the largest models, suggesting that gains from scaling do not straightforwardly translate into improved abstraction or socio-pragmatic sensitivity. Some of these task-level observations diverge from the ability-level ones, on the one hand underscoring the substantial variability in the ability categories that we have already mentioned and that is signaled by large error bars in the figures; on the other hand, a closer look at task-specific performance metrics might provide some explanation for such divergences and guide the development of more strongly harmonized metrics in the future. More task-specific analyses are provided in the next paragraph.

Multiple factors at play. While the detailed discussions in Section 4.1 highlight individual patterns across tasks, we conclude by drawing broader observations that emerge from the benchmark as a whole. In particular, CALAMITA allows us to disentangle how multiple, often intertwined factors, namely task design, domain specificity, and model scale, shape LLM performance in Italian.

A first clear pattern concerns tasks such as BEEP, INVALSI, and MULT-IT, which can be considered largely solved by the largest models, with LLAMA3.1-70B achieving around 80% accuracy. These challenges are grounded in factual knowledge, but what seems to play a decisive role is their multiple-choice formulation. This setup constrains the space of possible outputs and aligns well with models' instruction-tuning regimes and evaluation pipelines (e.g., `lm-eval-harness`), thereby minimizing error propagation due to generation or parsing. By contrast, other factual challenges, such as DIMMI and ECWCA, show substantially lower scores despite belonging to the same ability class, likely due to their open-ended structure and, in the case of DIMMI, their requirement for highly domain-specific medical knowledge. A second observation is that the apparent success of models on some tasks does not necessarily translate into robust generalization. For instance, even when LLMs excel in structured formats like

multiple choice, they often struggle with equivalent generative or reasoning formulations probing the same underlying competence. Overall, these results indicate that factors such as problem framing, domain familiarity, and response format can strongly modulate performance, sometimes overshadowing the effect of model size or linguistic exposure. Future work will leverage the richness of the CALAMITA benchmark, including its task descriptors and *Gist* files, to systematically quantify the impact of such variables, advancing our understanding of what LLMs truly learn versus what they adapt to through format and prompting.

Looking at the other end of the performance spectrum we find challenges like BLM-it and EurekaRebus, which remain very difficult even for the largest of the models we test. ECWCA also proves difficult. These challenges test the models' *linguistic* abilities, and so do TRACE-it and MACID, for example, for which scores are higher, especially for the former which results in an accuracy of above 60% for any model. However, BLM-it, EurekaRebus, and ECWCA are all framed as *puzzles*, introducing an additional layer of complexity. This suggests that when linguistic competence is embedded in playful or inferential reasoning settings, models still struggle to generalize beyond surface-level patterns. Adding multiple characterizations and levels of analysis might help to run deeper and more interpretable evaluations of the models' abilities in the future.

Last words on analysis. While overall we observe some clear trends, substantial differences across tasks, especially in the metrics used, but also in their setups and prompting strategies, prevent us from drawing broad generalizations (see also Section 6). A key future objective of the CALAMITA initiative is to promote greater standardization in the design of challenges and the definition of evaluation metrics, thereby facilitating more reliable aggregation and enabling clearer, higher-level insights.

5. Discussion and Outlook

Building on the results in Section 4, this section distills what we learned from coordinating CALAMITA as a community-run evaluation effort. Rather than revisiting per-task scores or aggregate scores, we focus on the broader insights that the process surfaced. After a first brief discussion on language-level takeaways that derive from working with and on Italian, and the picture that we can glean by looking at the challenges as a whole, rather than single scores, we expand on why CALAMITA is best viewed not as a competition among LLMs but as a *framework* that has the advantage of centralizing evaluation while distributing task development (Section 5.1). Based on these learned lessons, we elaborate on how we see future developments of this initiative (Section 5.2), reflecting on sustainability through rolling cycles, interaction with other evaluation initiatives, and the extent to which our approach yields a transferable blueprint for other communities.

5.1 Lessons

Native data and the case of Italian. Insisting on Italian-native datasets was crucial to focus on the socio-cultural and linguistic landscape of Italian, to avoid artifacts introduced by machine translation, and also to reduce training-set leakage due to replicas of long-existing datasets.

Besides the specific interest in promoting progress in Italian NLP, several properties make Italian a productive testbed: rich inflectional morphology (gender/number agreement), pro-drop with topic-comment organization, clitics, flexible word order, and

pervasive register variation. Especially in the context of the linguistic-based challenges, it is indeed along those points that one can interpret errors: agreement over long dependencies, role assignment in relative clauses, clitic-sensitive alternations, and genre-driven style constraints (see Section 4 for detailed discussions on these points). Due to linguistic differences, many of these aspects, and others such as those related to fairness in machine translation, for example, are muted in English-focused benchmarks.

Benchmarking as a distributed and centralized collaboration. Centralizing evaluation runs (common hardware, common evaluation settings and platforms, and centralized communication) while distributing task creation and development has proved to strike an excellent balance between inclusivity and diversity on the one hand, and consistency and standardization on the other. It reduces spurious variance due to unsystematic execution settings, thereby making cross-task comparison more sound, and at the same time, it allows task proposers from diverse, even non-technical fields to focus on a diverse range of linguistically and culturally meaningful challenges. Indeed, casting a wide net with a broadly distributed call for challenges and a light, inclusive pre-proposal screening coupled with hands-on support from the evaluation team lowered entry costs for highly diverse contributors. Shared templates (for reports and evaluation descriptors) also helped to uniform contributions at all stages of the pipeline.

From one-off campaign to community framework. The central lesson of CALAMITA is organizational rather than competitive. The initiative shows that a benchmarking effort can be operated as a *community service* with shared methods, artifacts, and governance rather than as a leaderboard race. The emphasis is not on “who wins" a task, but on how a research association can mobilize distributed expertise to (i) curate native, linguistically sound evaluation data, (ii) run fair and reproducible evaluations at scale, and (iii) return artifacts (scores, logs, and model outputs) that enable scientific interpretation. In this sense, CALAMITA is best understood as a *framework* for systematically probing LLM abilities in Italian.

5.2 Outlook

Following up on the observations in the previous section, we outline here what we plan for the future of CALAMITA.

Rolling process for continuous benchmarking. After the success and especially the community response of this first edition, our goal is to position CALAMITA as a permanent fixture within the Italian NLP research landscape to ensure the long-term sustainability of Italian benchmarking, and foster community engagement. We operationalize this through a *rolling process* that enables the continuous proposal and evaluation of new challenges designed to test specific abilities of LLMs. The technical and collaborative infrastructure that we have put in place to date will support this process, which will broadly follow the same pipeline (see Figure 1). As in the first round, we follow a three-stage process, comprising a pre-proposal screening, data and evaluation delivery, and submission of a short peer-reviewed paper. Once a proposal is accepted, we work with the challenge proposers to ensure their datasets and evaluation scripts are compatible with the lm-eval-harness framework, allowing our technical team to conduct standardized evaluations across multiple LLMs. Accepted papers are first published on arXiv and later included in the annual AILC proceedings, with results presented at CLiC-it or EVALITA. As in the first iteration, we prioritize datasets natively in Italian, ensure

reproducibility, and encourage the inclusion of private test splits to preserve benchmark integrity. Our first pilot cycle, targeting EVALITA 2026, runs from November 2025 to February 2026.

Running benchmarking in recurring cycles turns it into an ongoing process with regular stages for screening, feedback, evaluation, and publication. It also supports continuous monitoring of new developments and emerging research areas, including internationally. By supporting publication for the challenge papers and reserving space for discussion at regular meetings of the Italian Association for Computational Linguistics (CLiC-it and EVALITA) it also helps maintain community engagement and awareness of ongoing developments.

Model selection criteria. As future CALAMITA rounds will include the evaluation of additional models, an important next step will be to define more explicit and motivated criteria for model selection. In the present work, the evaluated models were chosen as representative open-weight systems that allowed us to test different dimensions of interest, such as model size, Italian-specific adaptation, and training from scratch on Italian data. As the benchmark grows, however, and new models are regularly released, model selection should increasingly support an architectural and explanatory analysis of model capabilities.

Future evaluations should make it possible to study how performance varies with scale, by comparing models from the same family but with different parameter counts, as well as models of similar size but different architectures, training-data composition, or adaptation strategies. At the same time, model selection should be considered together with task selection. CALAMITA should not only assess performance on application-oriented tasks, but also probe specific linguistic and cognitive abilities in Italian, including morphological analysis, agreement, syntactic disambiguation, lexical semantics, reasoning, and discourse-level interpretation. In this sense, future rounds should be designed to clarify not only which models perform best, but also which abilities they acquire, under which architectural and training conditions, and at what computational cost. Compute availability will also play a role in model selection.

Private datasets. Under the tenet that maximal reproducibility and open data are key to progress in science, all the datasets in the CALAMITA benchmark which are copyright-free, or in any case shareable, are accessible. They are on HuggingFace as well as accessible on the CALAMITA github[22]. While the fact that most datasets were newly created and even for previously existing ones not used in open evaluations (with the exception of part of ITAEVAL) overcomes the risk of them being included in the pre-training data of the LLMs we tested, the fact that they are openly accessible might pose this problem for newly trained models.[23] To balance the tension between open data and data leakage, thinking along the lines of Jacovi et al. (2023)'s advice, from the next round of CALAMITA's proposals onwards, it will be requested that as part of the data creation process two splits of the datasets are produced: a public one, which will be uploaded to HuggingFace, as it has been done for the previous iteration, and another, smaller one which will be sent directly to the CALAMITA board and kept private. Evaluations will be

22 https://github.com/calamita-AILC/calamita-eval

23 As an exception, the creators of TERMITE have ensured that their dataset is invisible to search engines thanks to an encryption key associated with the resource.

run on both splits, thereby also allowing for the influence of data leakages to be more readily detected.

Community building and mentoring lab. We also consider CALAMITA to be a prime opportunity for young researchers to gain experience through a collaborative effort of this size, learning technical ropes (running models on the supercomputer, fine-grained data processing, evaluation metrics), and coordination skills. The community atmosphere this endeavor fosters also promotes a healthy research attitude for future generations. To this end, we invite master's and early PhD students to join the evaluation team, and more senior PhD students and postdocs to join coordination activities.

CALAMITA *and* EVALITA. EVALITA, the evaluation campaign for Italian, which has reached its ninth edition, has fostered, in the course of over 18 years, the development of a vast number of datasets and systems, and the creation of an active community with collaborations across academia, industry, and NGOs (Basile et al. 2017; Passaro et al. 2020; Basile et al. 2022). Therefore, we care to elaborate briefly on the different roles of CALAMITA and EVALITA within the Italian NLP landscape. While both aim to advance the evaluation and development of language technologies for Italian, they do so through distinct yet potentially convergent approaches.

In the context of international benchmarking, CALAMITA aims to provide an overview of the capabilities of Italian language models (strictly native, not translated) through a range of tasks that together constitute a dynamic reference benchmark. The challenges included can be highly specific, covering multiple dimensions, also driven by curiosities emerging in fields other than NLP about LLMs' abilities. This initiative is crucial for proper, continuous Italian benchmarking.

EVALITA, on the other hand, continues to serve as an incentive to develop systems for concrete (and ideally useful and practical) tasks, by creating models that are not merely generic applications of an arbitrary LLM. The goal is to achieve the best possible performance through dedicated models, which are developed and fine-tuned specifically to solve a given task. The resulting systems can be distributed, further refined, and practically deployed for the specific purposes for which they were designed. Large language models may serve as baselines, but certainly not as definitive solutions.

The optimal scenario would be to integrate the two initiatives, and this is the approach the community would like to take in the future. One option is to identify which CALAMITA tasks lend themselves to an EVALITA-style approach, thereby fostering the development of dedicated models and maximizing the use of shared resources and synergies. Seen from the other perspective, all of the EVALITA past and future datasets could be reframed and integrated into the CALAMITA benchmark, such that LLM performance could be taken as a baseline for the EVALITA tasks. A few past EVALITA tasks are included in ITAEVAL, so there is already a good start to this process.

To conclude: by documenting this journey and how we plan to continue it, we aim to provide not only new empirical evidence on LLMs' competences in Italian, but also a practical blueprint for scalable benchmarking in less represented languages, which is grounded in association-backed governance, native data, standardized descriptors, centralized execution, and a rolling process to ensure dynamic growth.

Though rooted in Italian and AILC's long-standing tradition in evaluation, none of these elements is language-specific. More broadly, the value lies less in identifying winning models and more in helping the community ask better questions which are linguistically and culturally grounded, and produce findings that are reproducible and

explanatory. While CALAMITA is a significant step forward, substantial work remains to establish comprehensive and shared evaluation practices for Italian and other low-resource languages. We invite the international computational linguistics community to join us in advancing this collective effort.

6. Limitations

While we believe CALAMITA is a unique contribution to the Italian NLP landscape, and a useful methodological blueprint to extend collaborative benchmarking to yet other languages and communities, we are aware of a series of limitations of this work. We outline them, together with ideas for addressing them in the future.

First, the models evaluated are not task-optimized: a range of parameter choices such as prompt formulation, decoding strategy, context-window size, and few-shot selection could potentially improve performance. This was not the focus of the current setup, which instead aims to provide a consistent and future-resilient testing framework. In the future, evaluations could incorporate systematic hyperparameter tuning and controlled ablation studies to better isolate the impact of such choices. Until now, this strategy has been prevented not only by time issues but also by computational constraints, which limited the number of repetitions and robustness checks that could be performed; increasing available compute or implementing efficient evaluation pipelines would enable more thorough robustness analyses.

Second, the models we tested are not the most recent. They were selected with an eye on their Italian competence as well as simply being representative examples to demonstrate the benchmark's structure and interpretative potential. Thanks to the dynamic nature of the CALAMITA initiative and the centralized evaluation infrastructure, future work will naturally extend to new models. Relatedly, we tested exclusively open models, leaving out closed-source systems that dominate most practical and commercial applications as well as daily use of non-NLP researchers and lay users. While their inclusion is currently constrained by licensing and access restrictions, the benchmark is designed to accommodate such evaluations as well; collaborations with industry partners could be explored in the future to extend coverage to proprietary systems.

Third, the existing tasks lack full standardization, especially regarding metrics, as they have emerged from a range of research groups and, in some instances, from communities outside of NLP. This heterogeneity contributes valuable diversity to the benchmark and the tested model abilities, but simultaneously prevents optimal comparability of performance metrics across models and datasets. In next iterations we will need to establish shared evaluation protocols which while still leaving freedom to the proposers to choose the most appropriate metrics for their challenges would allow for more uniform evaluation conditions.

Lastly, our proposed categorization and taxonomy of tasks, while serving as a useful organizing and interpretative principle, remains somewhat arbitrary and should be regarded as a dynamic structure subject to change. Subsequent versions of the benchmark, with the addition of new tasks and testing further models, could refine this taxonomy through community input and, most of all, empirical validation by means of further rounds of analysis of the results. Hopefully, this way it will better serve the purpose of facilitating our understanding of model capabilities at a more general level, and shape the nature of future challenges.

Acknowledgments

We gratefully acknowledge CINECA and the EuroHPC Joint Undertaking (EuroHPC JU) for providing computational resources on the LEONARDO supercomputer (hosted by CINECA, Italy) under ISCRA grant HP10C3RW9F and ISCRA Class C grant CALAMITA (HP10CKZDYT). Additional computational resources were provided by the Center for Information Technology of the University of Groningen (Hábrók HPC cluster) and by the University of Turin (HPC4AI cluster). We also acknowledge CINECA-ISCRA support via Class C project IsCb7_LLMEVAL (HP10CIO7T9).

This work was partially supported by: the Dutch Sectorplan for the Humanities ("Humane AI" theme); the project "HARMONIA" (PNRR M4-C2, I1.3 Partenariati Estesi, Cascade Call FAIR; CUP C63C22000770006; PE0000013; NextGenerationEU); the PNRR project FAIR - Future Artificial Intelligence Research (PE00000013), including Spoke 2 "Integrative AI, (CUP C63C22000770006), Spoke 6 "Symbiotic AI" (CUP H97G22000210007), Spoke 8 "Pervasive AI", and related cascade calls (including Spoke 5 "High-Quality AI" cascade call "R4MSES", CUP B53C22003980006); the European Innovation Council project "EMERGE" (Grant No. 101070918); the ICSC - Centro Nazionale di Ricerca in "High Performance Computing, Big Data and Quantum Computing", Spoke "Future HPC & Big Data" (European Union - NextGenerationEU); the Portuguese Recovery and Resilience Plan project C645008882-00000055 (Center for Responsible AI) and FCT/MECI contract UIDB/50008 (Instituto de Telecomunicações); the PNRR project ITSERR (CUP B53C22001770006); PRIN 2022EPTPJ9 (WEMB), funded by the Italian Ministry of University and Research (MUR); PRIN2022 PRIMA (Ref. Prot. n. 20224TPEYC; CUP J53D23005130001); the project "ADELE - Analytics for DEcision of LEgal cases" (Justice Programme, GA No. 101007420); the project ATTRACTOR, funded under the call "Iniziative propedeutiche alla presentazione di Progetti di Ricerca volti a promuovere e favorire la definizione e presentazione di progetti Horizon/ERC - 2022" of the University of Naples "L'Orientale"; the CHIST-ERA project ANTIDOTE (ANR, Call XAI 2019; Project-ANR-21-CHR4-0002); the Swiss National Science Foundation (SNF Advanced grant TMAG-1_209426); the GARR "Orio Carlini" scholarship 2023/24; the Dutch Research Council (NWO) project InDeep (NWA.1292.19.399) and NWO Talent Programme (VI.Vidi.221C.009); Horizon Europe grant agreement No. 101135798 (Meetween); and Università degli Studi di Brescia support within the actions of its Gender Equality Plan. This work was also funded by the 'Multilingual PerspectiveAware NLU' project in partnership with Amazon Science, and partially supported by: DeepR3 (TED2021-130295B-C31), Disargue (TED2021-130810B-C21), and DeepKnowledge (PID2021-127777OB-C21) funded by MCIN/AEI/10.13039/501100011033 (and, where applicable, European Union NextGenerationEU/PRTR and FEDER, EU), as well as the Ixa group A type research group (IT1570-22) and IKER-GAITU project 11:4711:23:410:23/0808 funded by the Basque Government. Finally, this work was partially supported by ERC-2021-STG-101039777 (ABSTRACTION), funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union (or, where applicable, the European Research Council Executive Agency); neither the European Union nor the granting authority can be held responsible for them.

The authors gratefully acknowledge AILC for bringing together the Italian NLP community and enabling the collaborative effort underlying the CALAMITA initiative.

References

Amrhein, Chantal, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

An, Aixiu, Chunyang Jiang, Maria A. Rodriguez, Vivi Nastase, and Paola Merlo. 2023. BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1363–1374, Dubrovnik, Croatia, May.

Attanasio, Giuseppe, Pierpaolo Basile, Federico Borazio, Danilo Croce, Maria Francis, Jacopo Gili, Elio Musacchio, Malvina Nissim, Viviana Patti, Matteo Rinaldi, and Daniel Scalena. 2024a. CALAMITA: Challenge the abilities of LAnguage models in ITAlian. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1054–1063, Pisa, Italy, December. CEUR Workshop Proceedings.

Attanasio, Giuseppe, Pieter Delobelle, Moreno La Quatra, Andrea Santilli, and Beatrice Savoldi. 2024b. Itaeval and tweetyita: A new extensive benchmark and efficiency-first language model for italian. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Attanasio, Giuseppe, Moreno La Quatra, Andrea Santilli, and Beatrice Savoldi. 2024c. Itaeval: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Attardi, Giuseppe, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. 2015. State of the art language technologies for italian: The EVALITA 2014 perspective. *Intelligenza Artificiale*, 9(1):43–61.

Austin, Jacob, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Bai, Ge, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024a. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand, August. Association for Computational Linguistics.

Bai, Guangji, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, et al. 2024b. Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*.

Bandarkar, Lucas, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand, August. Association for Computational Linguistics.

Barbieri, Francesco, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016*, volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Basile, Pierpaolo, Elio Musacchio, and Lucia Siciliani. 2024. ITA-SENSE - evaluate llms' ability for italian word SENSE disambiguation: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Basile, Pierpaolo, Viviana Patti, Francesco Cutugno, Malvina Nissim, and Rachele Sprugnoli. 2017. Evalita goes social: Tasks, data. *IJCoL. Italian Journal of Computational Linguistics*, 3(3-1):93–127.

Basile, Valerio, Cristina Bosco, Michael Fell, Viviana Patti, and Rossella Varvara. 2022. Italian NLP for everyone: Resources and models from EVALITA to the European language grid. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 174–180, Marseille, France, June. European Language Resources

Association.
Basile, Valerio, Silvia Casola, Simona Frenda, and Soda Marem Lo. 2024. PERSEID - perspectivist irony detection: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.
Baucells, Irene, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE, January. Association for Computational Linguistics.
Biondo, Nicoletta, Elena Pagliarini, Vincenzo Moscati, Luigi Rizzi, and Adriana Belletti. 2023. Features matter: the role of number and gender features during the online processing of subject- and object- relative clauses in italian. *Language, Cognition and Neuroscience*, 38(6):802–820.
Bisk, Yonatan, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.
Bolondi, Giorgio and Clelia Cascella. 2017. Somministrazione delle prove invalsi dal 2009 al 2015: un patrimonio d'informazioni tra evidenze psicometriche e didattiche. In *I dati INVALSI: uno strumento per la ricerca*. Franco Angeli, Milano, page 14.
Brunato, Dominique. 2024. Trace-it: Testing relative clauses comprehension through entailment in italian: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.
Cafagna, Michele, Lorenzo De Mattei, Davide Bacciu, and Malvina Nissim. 2019. Suitable doesn't mean attractive. human-based evaluation of automatically generated headlines. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.
Cao, Yong, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In Sunipa Dev, Vinodkumar Prabhakaran, David Ifeoluwa Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia, May. Association for Computational Linguistics.
Casola, Silvia, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. Multipico: Multilingual perspectivist irony corpus. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16008–16021. Association for Computational Linguistics.
Cettolo, Mauro, Andrea Piergentili, Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. MAGNET - machines generating translations: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.
Chen, Jiaqi, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Chen, Po-Heng and Yun-Nung Chen. 2024. Efficient unseen language adaptation for multilingual pre-trained language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18983–18994, Miami, Florida, USA, November. Association for Computational Linguistics.

Chen, Zhuang, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand, August. Association for Computational Linguistics.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November. Association for Computational Linguistics.

Cignarella, Alessandra Teresa, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on irony detection in italian tweets (ironita). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Clark, Peter, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Costanzo, Antonella and Marta Desimoni. 2017. Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using invalsi survey data. *Large-scale Assessments in Education*, 5.

Croce, Danilo, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In Chiara Ghidini, Bernardo Magnini, Andrea Passerini, and Paolo Traverso, editors, *AI*IA 2018 - Advances in Artificial Intelligence - XVIIth International Conference of the Italian Association for Artificial Intelligence, Trento, Italy, November 20-23, 2018, Proceedings*, volume 11298 of *Lecture Notes in Computer Science*, pages 389–402. Springer.

Dell'Orletta, Felice, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors. 2024. *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Du, Weihong, Wenrui Liao, Hongru Liang, and Wenqiang Lei. 2024. PAGED: A benchmark for procedural graphs extraction from documents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10829–10846, Bangkok, Thailand, August. Association for Computational Linguistics.

Dua, Dheeru, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dugan, Liam, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand, August. Association for Computational Linguistics.

Fan, Lizhou, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2024. NPHardEval: Dynamic benchmark on reasoning ability of large language models via complexity classes. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4092–4114, Bangkok, Thailand, August. Association for Computational Linguistics.

Fersini, Elisabetta, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: automatic misogyny identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Francis, Maria, Matteo Rinaldi, Jacopo Gili, Leonardo De Cosmo, Sandro Iannaccone, Malvina Nissim, and Viviana Patti. 2024. GATTINA - generation of titles for italian news articles: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Frenda, Simona, Andrea Piergentili, Beatrice Savoldi, Marco Madeddu, Martina Rosola, Silvia Casola, Chiara Ferrando, Viviana Patti, Matteo Negri, and Luisa Bentivogli. 2024. GFG - gender-fair generation: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Frohberg, Jörg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France, June. European Language Resources Association.

Gao, Leo, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness, v0.4.3. https://zenodo.org/records/12608602. Zenodo.

Gili, Jacopo, Lucia C. Passaro, and Tommaso Caselli. 2023. Check-it!: A corpus of expert fact-checked claims for italian. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, and Nicole Novielli, editors, *Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023*, volume 3596 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Gili, Jacopo, Viviana Patti, Lucia C. Passaro, and Tommaso Caselli. 2024. Veryfit - benchmark of fact-checked claims for italian: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Giordano, Luca and Maria Pia di Buono. 2024. Large language models as drug information providers for patients. In Dina Demner-Fushman, Sophia Ananiadou, Paul Thompson, and Brian Ondov, editors, *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 54–63, Torino, Italy, May. ELRA and ICCL.

González, José Ángel, Ian Borrego Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Ríos, Angelo Basile, and Marc Franco-Salvador. 2026. Iberbench: Llm evaluation on iberian languages. *Computer Speech & Language*, 96:101899.

Gordon, Andrew S. 2016. Commonsense interpretation of triangle behavior. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3719–3725. AAAI Press.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and Anirudh Goyal et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grundler, Giulia, Andrea Galassi, Piera Santin, Alessia Fidelangeli, Federico Galli, Elena Palmieri, Francesca Lagioia, Giovanni Sartor, and Paolo Torroni. 2024. AMELIA - argument

mining evaluation on legal documents in italian: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Grundler, Giulia, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in CJEU decisions on fiscal state aid. In Gabriella Lapesa, Jodi Schneider, Yohan Jo, and Sougata Saha, editors, *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157, Online and in Gyeongju, Republic of Korea, October. International Conference on Computational Linguistics.

Guo, Taicheng, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36 of *NIPS '23*, pages 59662–59688, Red Hook, NY, USA, December. Curran Associates Inc.

Hartvigsen, Thomas, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May. Association for Computational Linguistics.

Helwe, Chadi, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico, June. Association for Computational Linguistics.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations (ICLR 2021)*, Virtual Event, May. OpenReview.net.

Hessel, Jack, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada, July. Association for Computational Linguistics.

Hu, Jennifer, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada, July. Association for Computational Linguistics.

Hu, Jennifer, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230499.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, pages 4411–4421. JMLR.org, July.

Huber, Eva, Sebastian Sauppe, Arrate Isasi-Isasmendi, Ina Bornkessel-Schlesewsky, Paola Merlo, and Balthasar Bickel. 2024. Surprisal from language models can predict ERPs in processing predicate-argument structures only if enriched by an Agent Preference principle. *Neurobiology of language*, 5(1):167–200.

Ichino, Pietro. 2021. *L'ora desiata vola: guida al mondo del rebus per solutori (ancora) poco abili*. Bompiani, Milan.

Jacovi, Alon, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 4615–4634, Bangkok, Thailand, August. Association for Computational Linguistics.

Jacovi, Alon, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, December. Association for Computational Linguistics.

Jeretic, Paloma, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online, July. Association for Computational Linguistics.

Jiang, Chunyang, Giuseppe Samo, Vivi Nastase, and Paola Merlo. 2024a. BLM-It - Blackbird Language Matrices for Italian: A CALAMITA Challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jiang, Yuxin, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024b. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand, August. Association for Computational Linguistics.

Joshi, Mandar, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July. Association for Computational Linguistics.

Kasner, Zdeněk and Ondrej Dusek. 2024. Beyond traditional benchmarks: Analyzing behaviors of open LLMs on data-to-text generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12045–12072, Bangkok, Thailand, August. Association for Computational Linguistics.

Khan, Mohammad Abdullah Matin, M Saiful Bari, Xuan Long Do, Weishi Wang, Md Rizwan Parvez, and Shafiq Joty. 2024. XCodeEval: An execution-based large scale multilingual multitask benchmark for code understanding, generation, translation and retrieval. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6766–6805, Bangkok, Thailand, August. Association for Computational Linguistics.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November. Association for Computational Linguistics.

Kosinski, Michal. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Krifka, Manfred, Francis Pelletier, Gregory N. Carlson, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. 1995. *Genericity: An introduction*. University of Chicago Press, 01.

Krumdick, Michael, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. BizBench: A quantitative reasoning benchmark for business and finance. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8309–8332, Bangkok, Thailand, August. Association for Computational Linguistics.

Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural

questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Ladhak, Faisal, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 4034–4048, Online, November. Association for Computational Linguistics.

Lai, Guokun, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

Landro, Nicola, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5).

Le, Matthew, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China, November.

Li, Bryan, Jiaming Luo, Eleftheria Briakou, and Colin Cherry. 2025. Leveraging domain knowledge at inference time for LLM translation: Retrieval versus generation. In Weijia Shi, Wenhao Yu, Akari Asai, Meng Jiang, Greg Durrett, Hannaneh Hajishirzi, and Luke Zettlemoyer, editors, *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 91–106, Albuquerque, New Mexico, USA, May. Association for Computational Linguistics.

Li, Qintong, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand, August. Association for Computational Linguistics.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Liang, Xun, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Cheng Peng, Zhonghao Wang, and Haiying Deng. 2024. UHGEval: Benchmarking the hallucination of Chinese large language models via unconstrained generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5266–5293, Bangkok, Thailand, August. Association for Computational Linguistics.

Lin, Chin-Yew. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS-2004)*, page 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May. Association for Computational Linguistics.

Liu, Chen, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022a. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Liu, Emmy, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States, July. Association for Computational Linguistics.

Liu, Xiao, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AlignBench: Benchmarking Chinese alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11621–11640, Bangkok, Thailand, August. Association for Computational Linguistics.

Lu, Shuai, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, Online Conference, Canada, December.

Ma, Bolei, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria, July. Association for Computational Linguistics.

Magnini, Bernardo, Roberto Zanoli, Michele Resta, Martin Cimmino, Paolo Albano, Marco Madeddu, and Viviana Patti. 2025. A Leaderboard for Benchmarking LLMs on Italian. In Cristina Bosco, Elisabetta Jezek, Marco Polignano, and Manuela Sanguinetti, editors, *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 636–646, Cagliari, Italy, September. CEUR Workshop Proceedings.

Magooda, Ahmed, Alec Helyar, Kyle Jackson, David Sullivan, Chad Atalla, Emily Sheng, Dan Vann, Richard Edgar, Hamid Palangi, Roman Lutz, et al. 2023. A framework for automated measurement of responsible ai harms in generative ai applications. *arXiv preprint arXiv:2310.17750*.

Manna, Raffaele, Maria Pia di Buono, and Luca Giordano. 2024. DIMMI - drug information mining in italian: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mercorio, Fabio, Daniele Potertì, Antonio Serino, and Andrea Seveso. 2024. BEEP - best driver's license performer: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Merlo, Paola. 2023. Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Motivations and formal specifications. *arXiv preprint arXiv:2306.11444*.

Miola, Emanuele. 2020. *Che cos'è un rebus*. Carocci.

Mokhberian, Negar, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico, June. Association for Computational Linguistics.

Moneglia, Massimo, Susan Brown, Francesca Frontini, Gloria Gagliardi, Anas Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014. The IMAGACT visual ontology. an extendable multilingual infrastructure for the representation of lexical encoding of action. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik,*

Iceland, May 26-31, 2014, pages 3425–3432. European Language Resources Association (ELRA).

Montes-y-Gómez, Manuel, Francisco Rangel, Salud María Jiménez-Zafra, Marco Casavantes, Begoña Altuna, Miguel Ángel Álvarez-Carmona, Gemma Bel-Enguix, Luis Chiruzzo, Iker de la Iglesia, Hugo Jair Escalante, Miguel Ángel García Cumbreras, José Antonio García-Díaz, José Ángel González Barba, Roberto Labadie Tamayo, Salvador Lima, Pablo Moral, Flor Miriam Plaza del Arco, and Rafael Valencia-García, editors. 2023. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023)*, volume 3496 of *CEUR Workshop Proceedings*, Jaén, Spain, September. CEUR-WS.org.

Muti, Arianna. 2024. Pejorativity - in-context pejorative language disambiguation: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nastase, Vivi, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024a. Exploring Italian sentence embeddings properties through multi-tasking. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1–10, Pisa, Italy, December. CEUR Workshop Proceedings.

Nastase, Vivi, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024b. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1–13, Pisa, Italy, December. CEUR Workshop Proceedings.

Navigli, Roberto and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July. Association for Computational Linguistics.

Nozza, Debora, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, June. Association for Computational Linguistics.

Orlando, Riccardo, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy, December. CEUR Workshop Proceedings.

Paperno, Denis, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany, August. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Parmar, Mihir, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand, August. Association for Computational Linguistics.

Passaro, Lucia C., Maria Di Maro, Valerio Basile, and Danilo Croce. 2020. Lessons learned from evalita 2020 and thirteen years of evaluation of italian language technology. *IJCoL. Italian*

Journal of Computational Linguistics, 6(6-2):79–102.
Pensa, Giulia, Ekhi Azurmendi, Julen Etxaniz, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. GITA4CALAMITA - evaluating the physical commonsense understanding of italian llms in a multi-layered approach: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.
Piazzolla, Silvia Alma, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *HERMES - Journal of Language and Communication in Business*, (63):209–225, Dec.
Piergentili, Andrea, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore, December. Association for Computational Linguistics.
Piergentili, Andrea, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK, June. European Association for Machine Translation (EAMT).
Pietschnig, Jakob, Sandra Oberleiter, Enrico Toffalini, and David Giofrè. 2023. Reliability of the g factor over time in italian invalsi data (2010-2022): What can achievement-g tell us about the flynn effect? *Personality and Individual Differences*, 214:112345.
Plaza, Irene, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and LLM benchmarks: is MMLU lost in translation? *arXiv preprint arXiv:2406.17789*.
Polignano, Marco, Pierpaolo Basile, and Giovanni Semeraro. 2026. Advanced Natural-based interaction for the ITAlian language: LLaMAntino-3-ANITA. *Scientific Reports*, 16.
Ponti, Edoardo Maria, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online, November. Association for Computational Linguistics.
Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.
Puccetti, Giovanni, Maria Cassese, and Andrea Esuli. 2024. INVALSI - mathematical and language understanding in italian: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.
Puccetti, Giovanni, Maria Cassese, and Andrea Esuli. 2025. The invalsi benchmarks: measuring the linguistic and mathematical understanding of large language models in Italian. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6782–6797, Abu Dhabi, UAE, January. Association for Computational Linguistics.
Puccetti, Giovanni, Claudia Collacciani, Andrea Amelio Ravelli, Andrea Esuli, and Marianna Bolognesi. 2024. ABRICOT - abstractness and inclusiveness in context: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ranaldi, Federico, Elena Sofia Ruzzetti, Dario Onorati, Fabio Massimo Zanzotto, and Leonardo Ranaldi. 2024. Termite italian text-to-sql: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ravelli, Andrea Amelio, Rossella Varvara, and Lorenzo Gregori. 2024. MACID - multimodal action identification: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ravichander, Abhilasha, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China, November. Association for Computational Linguistics.

Rei, Ricardo, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA, November. Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Rinaldi, Matteo, Jacopo Gili, Maria Francis, Mattia Goffetti, Viviana Patti, and Malvina Nissim. 2024. Mult-it multiple choice questions on multiple topics in italian: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rizzi, Giulia, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italy, May. ELRA and ICCL.

Rohrbach, Anna, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, Hugo Larochelle, Aaron C. Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Rosola, Martina, Simona Frenda, Alessandra Teresa Cignarella, Matteo Pellegrini, Andrea Marra, and Mara Floris. 2023. Beyond obscuration and visibility: Thoughts on the different strategies of gender-fair language in italian. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, and Nicole Novielli, editors, *Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023*, volume 3596 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rosso, Paolo, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz, editors. 2018. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Röttger, Paul, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid), July. Association for Computational Linguistics.

Sabour, Sahand, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar,

editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand, August. Association for Computational Linguistics.

Sadat, Mobashir and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland, May. Association for Computational Linguistics.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, August.

Sakurai, Hiromasa and Yusuke Miyao. 2024. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657, Bangkok, Thailand, August.

Samo, Giuseppe, Vivi Nastase, Chunyang Jiang, and Paola Merlo. 2023. BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 12276–12287, Singapore, December. Association for Computational Linguistics.

Sanguinetti, Manuela, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2 @ EVALITA2020: overview of the EVALITA 2020 hate speech detection task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Sap, Maarten, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November. Association for Computational Linguistics.

Sarti, Gabriele, Tommaso Caselli, Arianna Bisazza, and Malvina Nissim. 2024a. Eurekarebus - verbalized rebus solving with llms: A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sarti, Gabriele, Tommaso Caselli, Malvina Nissim, and Arianna Bisazza. 2024b. Non verbis, sed rebus: Large language models are weak solvers of Italian rebuses. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 888–897, Pisa, Italy, December. CEUR Workshop Proceedings.

Sarvazyan, Areg Mikael, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural*, 71:275–288.

Savoldi, Beatrice, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore, December. Association for Computational Linguistics.

Shapira, Natalie. 2023. *Challenging Natural Language Processing through the Lens of Psychology*. Ph.D. thesis, Bar-Ilan University.

Shapira, Natalie, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta, March. Association for Computational Linguistics.

Siska, Charlotte, Katerina Marazopoulou, Melissa Ailem, and James Bono. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10406–10421, Bangkok, Thailand, August. Association for Computational Linguistics.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, May.

Stowe, Kevin, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland, May. Association for Computational Linguistics.

Suijkerbuijk, Michelle, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4):1267–1301, December.

Sun, Jiaxing, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking Chinese commonsense reasoning of LLMs: From Chinese-specifics to reasoning-memorization correlations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11205–11228, Bangkok, Thailand, August. Association for Computational Linguistics.

Tao, Yan, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):1–9, 09.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Thellmann, Klaudia, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

Todd, Graham, Tim Merino, Sam Earle, and Julian Togelius. 2024. Missed connections: Lateral thinking puzzles for large language models. In *2024 IEEE Conference on Games (CoG)*, pages 1–8, Milan, Italy, August. IEEE.

Tolosani, Demetrio. 1901. *Enimmistica*. Hoepli, Milan.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trotta, Daniela, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Tu, Shangqing, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. 2024. WaterBench: Towards holistic evaluation of watermarks for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1542, Bangkok, Thailand, August. Association for Computational Linguistics.

Turisini, Matteo, Giorgio Amati, and Mirko Cestari. 2024. LEONARDO: A pan-european pre-exascale supercomputer for HPC and AI applications. *Journal of large-scale research facilities*, 9(1).

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose

language understanding systems. *Advances in neural information processing systems*, 32.
Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.
Wang, Fali, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2025. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87.
Wang, Yan, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark, September. Association for Computational Linguistics.
Wang, Yuhao, Yusheng Liao, Heyang Liu, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. MM-SAP: A comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9192–9205, Bangkok, Thailand, August. Association for Computational Linguistics.
Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.
Welbl, Johannes, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, September. Association for Computational Linguistics.
Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.
Wu, Minghao, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *ArXiv*, abs/2504.15521.
Xia, Congying, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. FOFO: A benchmark to evaluate LLMs' format-following capability. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–699, Bangkok, Thailand, August. Association for Computational Linguistics.
Xu, Hainiu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand, August. Association for Computational Linguistics.
Xuan, Weihao, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China, November. Association for Computational Linguistics.
Yan, Weixiang, Haitian Liu, Yunkun Wang, Yunzhe Li, Qian Chen, Wen Wang, Tingyu Lin, Weishan Zhao, Li Zhu, Hari Sundaram, and Shuiguang Deng. 2024. CodeScope: An execution-based multilingual multitask multidimensional benchmark for evaluating LLMs on code understanding and generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar,

editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5511–5558, Bangkok, Thailand, August. Association for Computational Linguistics.

Yang, Yinfei, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China, November. Association for Computational Linguistics.

Yim, Wen-wai, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

Yu, Tao, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zaninello, Andrea, Sofia Brenna, and Bernardo Magnini. 2023. Textual entailment with natural language explanations: The italian e-rte-3 dataset. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, and Nicole Novielli, editors, *Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023*, volume 3596 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zaninello, Andrea and Bernardo Magnini. 2024. GEESE - generating and evaluating explanations for semantic entailment: a CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zaranis, Emmanouil, Giuseppe Attanasio, Sweta Agrawal, and André F. T. Martins. 2025. Watching the watchers: Exposing gender disparities in machine translation quality estimation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25261–25284, Vienna, Austria, July. Association for Computational Linguistics.

Zelikman, Eric, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. Quiet-STar: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, Philadelphia, PA, October.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July. Association for Computational Linguistics.

Zhang, Bin, Yuxiao Ye, Guoqing Du, Xiaoru Hu, Zhishuai Li, Sun Yang, Chi Harold Liu, Rui Zhao, Ziyue Li, and Hangyu Mao. 2024a. Benchmarking the text-to-sql capability of large language models: A comprehensive evaluation. *arXiv preprint arXiv:2403.02951*.

Zhang, Xinrong, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. ∞Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand, August. Association for Computational Linguistics.

Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024c. SafetyBench: Evaluating the safety of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand, August. Association for Computational Linguistics.

Zhang, Zhihan, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024d. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 1588–1606, Bangkok, Thailand, August. Association for Computational Linguistics.

Zhang, Ziyin, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024e. MELA: Multilingual evaluation of linguistic acceptability. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand, August. Association for Computational Linguistics.

Zheng, Jonathan, Alan Ritter, and Wei Xu. 2024. NEO-BENCH: Evaluating robustness of large language models with neologisms. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand, August. Association for Computational Linguistics.

Zheng, Zilong, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online, August. Association for Computational Linguistics.

Zhong, Ming, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online, June. Association for Computational Linguistics.

Zhong, Wanjun, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico, June. Association for Computational Linguistics.

Zugarini, Andrea, Kamyar Zeinalipour, Achille Fusco, and Asya Zanollo. 2024a. ECWCA - educational crossword clues answering A CALAMITA challenge. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zugarini, Andrea, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024b. Clue-instruct: Text-based clue generation for educational crossword puzzles. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 3347–3356. ELRA and ICCL.

Appendix A: Full Results

This appendix provides the complete quantitative results for all challenges. For each challenge we include the full set of subtasks and metrics. In addition, we detail the mapping between abilities, challenges, and (sub)tasks used to build Table 1 and to derive the aggregated views reported in Figures 8 and 9.

Table A.1
Summary of results across all CALAMITA challenges. Each row reports the models' average performance on the corresponding task.

Parent	Task	Metric	M	A	L8B	L70B
ABRICOT	abs	pearson	-0.03	0.50	0.29	0.44
	inc	pearson	-0.20	0.14	0.18	**0.25**
AMELIA	arg-component-fewshot	f1	0.23	0.41	0.48	**0.86**
	arg-component-zeroshot	f1	0.00	0.44	0.47	**0.71**
	arg-premisetype-fewshot	f1	0.48	0.60	0.74	**0.86**
	arg-premisetype-zeroshot	f1	0.57	0.57	0.39	**0.85**
	arg-scheme-fewshot	f1	0.27	0.28	0.42	**0.57**
	arg-scheme-zeroshot	f1	0.16	0.21	0.34	**0.42**
BEEP	beep	accuracy	0.52	0.62	0.65	**0.84**
BLM-It	agr1_0shots	f1	0.03	0.14	0.10	**0.33**
	agr1_1shots	f1	0.07	0.18	0.26	**0.41**
	agr2_0shots	f1	0.02	0.20	0.11	**0.35**
	agr2_1shots	f1	0.09	0.30	0.24	**0.41**
	caus1_0shots	f1	0.03	0.05	0.05	**0.09**
	caus1_1shots	f1	0.09	0.08	0.13	**0.18**
	caus2_0shots	f1	0.03	0.07	0.06	**0.08**
	caus2_1shots	f1	0.09	0.12	0.13	**0.19**
	od1_0shots	f1	0.03	0.06	0.06	**0.07**
	od1_1shots	f1	0.09	0.09	0.13	**0.27**
	od2_0shots	f1	0.03	0.06	0.06	**0.07**
	od2_1shots	f1	0.10	0.10	0.12	**0.20**
DIMMI	global	accuracy	0.07	**0.34**	0.31	**0.34**
	p1-molecule	accuracy	0.01	0.64	0.86	**0.87**
	p1-usage	accuracy	0.08	0.12	**0.19**	**0.19**
	p1-drug_interaction	accuracy	0.12	0.21	**0.43**	0.39
	p1-posology	accuracy	0.17	0.16	0.17	**0.18**
	p1-side_effect	accuracy	0.01	0.07	**0.16**	0.15
ECWCA	hint	f1	0.52	0.10	0.42	**0.67**
	no-hint	f1	0.54	0.08	0.43	**0.66**
EUREKAREBUS	eureka_hints	accuracy	0.00	0.00	0.07	**0.32**
	eureka_original	accuracy	0.00	0.00	0.10	**0.36**
GATTINA	ansa	sbert_score	0.07	0.17	0.33	**0.59**
	galileo	sbert_score	0.21	0.21	0.22	**0.26**
GEESE	anon_anita	accuracy	0.57	0.83	0.66	**0.89**
	anon_dummy	accuracy	0.49	0.54	0.49	**0.61**
	anon_gold	accuracy	0.58	0.71	0.62	**0.82**
	anon_llama	accuracy	0.57	0.79	0.60	**0.86**
	noexp	accuracy	0.49	0.47	0.50	**0.57**
GFG	task_1_1	bert_f1	0.49	0.55	0.66	**0.59**
	task_1_2	bert_f1	0.17	0.45	0.50	**0.53**
	task_2_1	acc_gente	0.53	0.51	0.18	**0.61**
	task_2_2	cwa	0.45	0.54	0.53	**0.73**
	task_2_3	acc_gente	0.33	0.50	0.21	**0.54**
	task_3_1	cwa	0.28	0.35	0.42	**0.58**
	task_3_2	acc_gente	0.59	0.57	**0.61**	0.56

Table A.1 continued from previous page

Parent	Task	Metric	M	A	L8B	L70B
GITA4CALAMITA	conflict_consistency	accuracy	0.02	0.18	0.29	**0.65**
	physical_verifiability	accuracy	0.00	0.08	0.14	**0.36**
	story_class_accuracy	accuracy	0.38	0.59	0.72	**0.88**
INVALSI	ita	accuracy	0.38	0.71	0.71	**0.89**
	ita_binarie	accuracy	0.59	0.65	0.61	**0.74**
	ita_multipla	accuracy	0.35	0.72	0.73	**0.91**
	mate	accuracy	0.34	0.47	0.51	**0.72**
	mate_multipla	accuracy	0.30	0.41	0.45	**0.70**
	mate_numero	accuracy	0.27	0.54	0.59	**0.78**
	mate_verofalso	accuracy	0.59	0.61	0.59	**0.65**
ITA-SENSE	gen-no-translation	rougeBertScore	0.26	0.26	**0.32**	0.31
	gen-with-translation	rougeBertScore	0.25	0.26	**0.32**	0.31
	ml-no-translation	extract_answer	0.22	0.51	0.41	**0.63**
	ml-with-translation	extract_answer	0.20	0.48	0.39	**0.58**
ITAEVAL	ami_2020_aggressiv.	f1	0.44	0.48	**0.60**	0.52
	ami_2020_misogyny	f1	0.52	0.73	0.73	**0.85**
	gente_rephrasing	accuracy	0.26	0.35	0.31	**0.43**
	haspeede2_hs	f1	0.52	0.70	0.70	**0.73**
	haspeede2_stereo	f1	0.46	0.62	0.60	**0.67**
	hatecheck_ita	f1	0.71	0.81	0.83	**0.89**
	honest_ita	accuracy	**1.00**	**1.00**	**1.00**	**1.00**
	ironita_irony	f1	0.42	0.69	0.67	**0.79**
	ironita_sarcasm	f1	0.42	0.46	0.52	**0.61**
	itacola	accuracy	0.74	0.69	0.82	**0.88**
	news_sum_fanpage	rouge1	0.29	0.30	**0.33**	0.31
	news_sum_ilpost	rouge1	0.28	0.29	**0.32**	0.27
	sentipolc	f1	0.44	0.50	0.49	**0.57**
MACID	macid	accuracy	0.27	0.43	0.43	**0.58**
MAGNET	en_it_public	bleu	0.28	0.25	0.27	**0.32**
	IT_en_it_private	bleu	0.43	0.35	0.41	**0.51**
	it_en_public	bleu	0.33	0.31	0.35	**0.38**
	IT_it_en_private	bleu	0.47	0.33	0.47	**0.53**
	UK_en_it_private	bleu	0.42	0.31	0.40	**0.50**
	UK_it_en_private	bleu	0.47	0.32	0.49	**0.54**
	US_en_it_private	bleu	0.33	0.24	0.30	**0.34**
	US_it_en_private	bleu	0.37	0.26	0.40	**0.43**
MULTI-It	multi-it-a	accuracy	0.39	0.62	0.65	**0.84**
	multi-it-c	accuracy	0.50	0.63	0.66	**0.81**
PEJORATIVITY	misogyny	accuracy	0.61	0.67	0.46	**0.75**
	misogyny-context	accuracy	0.66	0.79	**0.82**	0.80
	standard	accuracy	0.43	0.51	0.44	**0.59**
PERSEID	task_0	f1	0.32	0.34	0.49	**0.50**
	task_1	f1	0.38	0.29	0.50	**0.50**
	task_2	f1	0.35	0.31	0.50	**0.50**
	task_3	f1	0.39	0.23	0.50	**0.49**
TERMite	ita-text-to-sql	exec_accuracy	0.04	0.38	0.36	**0.46**
TRACE-IT	traceIT	accuracy	0.63	0.70	0.72	**0.85**
VERYf-IT	enriched	accuracy	**0.57**	0.43	0.52	0.56
	full	accuracy	**0.59**	0.41	0.52	**0.59**
	small	accuracy	**0.57**	0.43	0.52	0.56

Table A.2
Tasks grouped by category, with counts and lists of top-level tasks and subtasks.

Ability	#	Tasks
Common.	10	AMELIA, ECWCA, EUREKAREBUS, GEESE, GITA4CALAMITA, INVALSI, ITAEVAL, MACID, MULTI-It, PERSEID
	42	amelia-arg-component-fewshot, amelia-arg-component-zeroshot, amelia-arg-premisetype-fewshot, amelia-arg-premisetype-zeroshot, amelia-arg-scheme-fewshot, amelia-arg-scheme-zeroshot, ami_2020_aggressiveness, ami_2020_misogyny, conflict_consistency, ecwca-hint, ecwca-no-hint, eureka_hints, eureka_original, geese_anon_anita, geese_anon_dummy, geese_anon_gold, geese_anon_llama, geese_noexp, gente_rephrasing, haspeede2_hs, haspeede2_stereo, hatecheck_ita, honest_ita, invalsi_ita, invalsi_ita_binarie, invalsi_ita_multipla, invalsi_mate, invalsi_mate_multipla, invalsi_mate_numero, invalsi_mate_verofalso, ironita_irony, ironita_sarcasm, macid, multi-it-a, multi-it-c, perse_task_0, perse_task_1, perse_task_2, perse_task_3, physical_verifiability, sentipolc, story_class_accuracy
Factual	8	AMELIA, BEEP, DIMMI, ECWCA, EUREKAREBUS, INVALSI, MULTI-It, VERYFIT
	29	amelia-arg-component-fewshot, amelia-arg-component-zeroshot, amelia-arg-premisetype-fewshot, amelia-arg-premisetype-zeroshot, amelia-arg-scheme-fewshot, amelia-arg-scheme-zeroshot, beep, dimmi-global, dimmi-p1-drug_interaction, dimmi-p1-molecule, dimmi-p1-posology, dimmi-p1-side_effect, dimmi-p1-usage,dimmi-p2-drug_interaction, dimmi-p2-molecule, dimmi-p2-posology, dimmi-p2-side_effect, dimmi-p2-usage, ecwca-hint, ecwca-no-hint, eureka_hints, eureka_original, invalsi_ita, invalsi_ita_binarie, invalsi_ita_multipla, invalsi_mate, invalsi_mate_multipla, invalsi_mate_numero, invalsi_mate_verofalso, multi-it-a, multi-it-c, veryfIT_enriched, veryfIT_full, veryfIT_small
Linguistic	14	ABRICOT, BLM-It, ECWCA, EUREKAREBUS, GFG, INVALSI, ITA-SENSE, ITAEVAL, MACID, MULTI-It, PEJORATIVITY, PERSEID, TRACE-IT, VERYFIT
	60	abricot_abs, abricot_inc, ami_2020_aggressiveness, ami_2020_misogyny, blm_agr1_0shots, blm_agr1_1shots, blm_agr2_0shots, blm_agr2_1shots, blm_caus1_0shots, blm_caus1_1shots, blm_caus2_0shots, blm_caus2_1shots, blm_od1_0shots, blm_od1_1shots, blm_od2_0shots, blm_od2_1shots, ecwca-hint, ecwca-no-hint, eureka_hints, eureka_original, gente_rephrasing, gfg_task_1_1, gfg_task_1_2, gfg_task_2_1, gfg_task_2_2, gfg_task_2_3, gfg_task_3_1, gfg_task_3_2, haspeede2_hs, haspeede2_stereo, hatecheck_ita, invalsi_ita, invalsi_ita_binarie, invalsi_ita_multipla, invalsi_mate, invalsi_mate_multipla, invalsi_mate_numero, invalsi_mate_verofalso, ironita_irony, ironita_sarcasm, ita-sense-gen-no-translation, ita-sense-gen-with-translation, ita-sense-ml-no-translation, ita-sense-ml-with-translation, itacola, macid, multi-it-a, multi-it-c, pejorativITy-misogyny, pejorativITy-misogyny-context, pejorativITy-standard, perse_task_0, perse_task_1, perse_task_2, perse_task_3, sentipolc, traceIT, veryfIT_enriched, veryfIT_full, veryfIT_small
Reasoning	7	ECWCA, EUREKAREBUS, GEESE, GITA4CALAMITA, INVALSI, MULTI-It, TRACE-IT
	22	conflict_consistency, ecwca-hint, ecwca-no-hint, eureka_hints, eureka_original, geese_anon_anita, geese_anon_dummy, geese_anon_gold, geese_anon_llama, geese_noexp, invalsi_ita, invalsi_ita_binarie, invalsi_ita_multipla, invalsi_mate, invalsi_mate_multipla, invalsi_mate_numero, invalsi_mate_verofalso, multi-it-a, multi-it-c, physical_verifiability, story_class_accuracy, traceIT
Fairness	3	GFG, ITAEVAL, PEJORATIVITY
	16	ami_2020_aggressiveness, ami_2020_misogyny, gente_rephrasing, gfg_task_1_1, gfg_task_1_2, gfg_task_2_1, gfg_task_2_2, gfg_task_2_3, gfg_task_3_1, gfg_task_3_2, haspeede2_hs, haspeede2_stereo, hatecheck_ita, pejorativITy-misogyny, pejorativITy-misogyny-context, pejorativITy-standard
Code	1	TERMITE
	1	ita-text-to-sql
MT	2	MAGNET, GFG
	15	MAGNET_IT_en_it_private, MAGNET_IT_it_en_private, MAGNET_UK_en_it_private, MAGNET_UK_it_en_private, MAGNET_US_en_it_private, MAGNET_US_it_en_private, MAGNET_en_it_public, MAGNET_it_en_public, gfg_task_1_1, gfg_task_1_2, gfg_task_2_1, gfg_task_2_2, gfg_task_2_3, gfg_task_3_1, gfg_task_3_2
Summar.	2	GATTINA, ITAEVAL
	4	gattina-ansa, gattina-galileo, news_sum_fanpage, news_sum_ilpost

Table A.3
Challenges and associated abilities.

Task	Commonsense	Factual	Linguistic	Reasoning	Fairness	Code	MT	Summariz.
ABRICOT			✓					
AMELIA	✓	✓						
BEEP		✓						
BLM			✓					
DIMMI		✓						
ECWCA	✓	✓	✓	✓				
EUREKAREBUS	✓	✓	✓	✓				
GATTINA								✓
GEESE	✓			✓				
GFG			✓		✓		✓	
GITA4CALAMITA	✓			✓				
INVALSI	✓	✓	✓	✓				
ITA-SENSE			✓					
ITAEVAL	✓		✓		✓			✓
MACID	✓		✓					
MAGNET							✓	
MULT-IT	✓	✓	✓	✓				
PEJORATIVITY			✓		✓			
PERSEID	✓		✓					
TERMITE						✓		
TRACEIT			✓	✓				
VERYFIT		✓	✓					

Table A.4
Results on the six ITAEVAL tasks which have been translated from English and are not part of the official CALAMITA benchmark; see Section 4, and in particular Section 4.1.22 for details.

Task	**Metric**	M	A	L8B	L70B
arc_challenge_ita	accuracy	0.37	**0.53**	0.41	**0.53**
belebele_ita	accuracy	0.42	0.84	0.86	**0.92**
hellaswag_ita	accuracy	0.45	0.51	0.45	**0.53**
squad_it	squad_em	0.00	0.53	0.65	**0.66**
truthfulqa_mc2_ita	accuracy	0.41	**0.68**	0.51	0.54
xcopa_it	accuracy	0.74	0.74	0.72	**0.82**