INVALSI - Mathematical and Language Understanding in Italian: A CALAMITA Challenge

Giovanni Puccetti[1,*], Maria Cassese[1] and Andrea Esuli[1]

[1]Istituto di Scienza e Tecnologia dell'Informazione - CNR

Abstract

While Italian is a high resource language, there are few Italian-native benchmarks to evaluate Language Models (LMs) generative abilities in this language. This work presents two new benchmarks: Invalsi MATE to evaluate models performance on mathematical understanding in Italian and Invalsi ITA to evaluate language understanding in Italian.

These benchmarks are based on the Invalsi tests, which are administered to students of age between 6 and 18 within the Italian school system. These tests are prepared by expert pedagogists and have the explicit goal of testing average students' performance over time across Italy. Therefore, the questions are well written, appropriate for the age of the students, and are developed with the goal of assessing students' skills that are essential in the learning process, ensuring that the benchmark proposed here measures key knowledge for undergraduate students.

Invalsi MATE is composed of 420 questions about mathematical understanding, these questions range from simple money counting problems to Cartesian geometry questions, e.g. determining if a point belongs to a given line. They are divided into 4 different types: *scelta multipla* (multiple choice), *vero/falso* (true/false), *numero* (number), *completa frase* (fill the gap).

Invalsi ITA is composed of 1279 questions regarding language understanding, these questions involve both the ability to extract information and answer questions about a text passage as well as questions about grammatical knowledge. They are divided into 4 different types: *scelta multipla* (multiple choice), *binaria* (binary), *domanda aperta* (open question), *altro* (other).

We evaluate 4 powerful language models both English-first and tuned for Italian to see that best accuracy on Invalsi MATE is 55% while best accuracy on Invalsi ITA is 80%.

Keywords

Mathematical Understanding, Language Understanding, Invalsi, Large Language Models, Italian Language Models

1. Challenge: Introduction and Motivation

Assessing the quality of Large Language Models is a challenging task because these models can virtually perform any task that can be presented through natural language. To address this difficulty, each model needs to be tested on several tasks at once.

To help provide new benchmarks to evaluate LLMs in Italian, We propose two benchmarks, Invalsi MATE and Invalsi ITA the first meant to evaluate LLMs' mathematical understanding and the second to evaluate their language understanding, both in Italian.

These benchmarks originate from the Invalsi tests, which have been used in the past for demographic studies [1, 2, 3], but, to the best of our knowledge we are the first to use them to test LLMs performance in Italian [4], followed only later by others [5].

There are several benchmarks to evaluate mathematical understanding of LLMs based on English tests [6, 7, 8] and there are also several multi-domain benchmarks involving Italian [9], however there aren't any specifically focused on mathematical understanding in Italian. We focus on high-school questions, an English benchmark similar to Invalsi MATE is the GSM8k one, [8], which contains 8,500 high-school questions.

Language Models understanding of language in English is also well studied, there are several benchmarks meant to measure the ability of language models to understand language constructs in English, such as [10, 11], also arranged into extensive suites [12]. On the contrary there are fewer examples of these tests for the Italian language.

Therefore we propose Invalsi ITA which contains questions that are usually split among several different benchmarks, e.g. MNLI [13], SQuAD [14] and others from the GLUE suite [15]. The questions in the dataset cover several aspects of language understanding, ranging from the ability to extract specific information, such as the date when something happened to more complex information such as whether two events implicate each other or not.

These two datasets allow us to measure two key abilities of language models in Italian, to make the comparison among different models more fair we cast all questions as multiple choice and measure models' performance by selecting the answer with the highest likelihood according

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

[*]Corresponding author.

giovanni.puccetti@isti.cnr.it (G. Puccetti); maria.cassese@isti.cnr.it (M. Cassese); andrea.esuli@isti.cnr.it (A. Esuli)

https://gpucce.github.io/ (G. Puccetti); https://www.isti.cnr.it/en/about/people-detail/1072/Maria_Cassese (M. Cassese); https://www.esuli.it/ (A. Esuli)

0000-0002-5725-4322 (A. Esuli)

Testo
Elisa è uscita da casa questa mattina alle ore 8:15. Elisa è rientrata nel pomeriggio alle ore 1:15
Domanda
Quanto tempo è stata fuori casa Elisa?
A. 5 ore B. 7 ore C. 9 ore D. 11 ore

(a) *scelta multipla* question from Invalsi MATE.

Testo
Se moltiplichi per 2 un numero naturale e dal risultato sottrai 1, ottieni sempre un numero pari.
Domanda
Vero o Falso?

(b) *vero/falso* question from Invalsi MATE.

Testo
Filippo dice: per trovare il numero della mia maglietta aggiungi una decina e sei unità al numero 4.
Domanda
Qual è il numero della maglietta di Filippo?

(c) *numero* question from Invalsi MATE.

Testo
Luca lancia due dadi a sei facce non truccati.
Domanda
Completa la frase inserendo una delle espressioni:
La probabilità che la somma dei punti sia 12 è *maggiore della, minore della, uguale alla* probabilità che la somma sia 2.

(d) *completa frase* question from Invalsi MATE.

Figure 1: Examples of each question type from the Invalsi MATE dataset.

to the model.

We measure the performance of 4 strong large models, *mixtral instruct* [16], *mistral instruct* [17], *llama 3 8b instruct* [18], *anita 8b dpo* [19], the first three are English-first and the fourth is fine-tuned in Italian, and the current only Italian-first model *minerva 3b*. We show that on Invalsi ITA the best model among those we tested is *mixtral instruct*, which reaches an accuracy of 0.8, while on Invalsi MATE the highest accuracy is 0.55, also achieved by *mixtral instruct*.

Both language understanding and mathematical understanding are key abilities for students as well as language models, particularly since these models are often used in learning environments. By adding these benchmarks to the CALAMITA suite we hope they will help the development of LLMs in Italian by providing a more comprehensive evaluation of their abilities and thus fostering the research and development of models in this language.

The CALAMITA special event [20], which has aims to establishing a shared benchmark for LLMs in Italian, is a first step towars a systematic evaluation of LLMs in this language. We hope that the Invalsi challenge will enrich the Linguistic and Mathematical understanding branches of this shared benchmark.

2. Challenge: Description

The challenge is composed of two tasks: Invalsi MATE and Invalsi ITA. For each task, we provide a detailed description of the data, the metrics used for evaluation, and the limitations of the data.

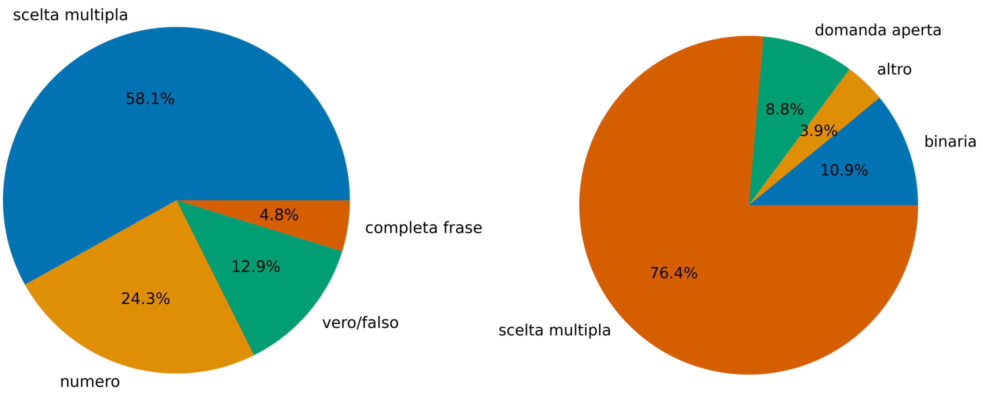
2.1. Task 1: Mathematical Understanding in Italian (Invalsi MATE)

The first task consists in answering mathematical questions in Italian. These questions are meant for students from 6 to 18 years of age, therefore the kind of question can vary significantly, from simpler, example-based, ones that don't require any knowledge besides counting, to more complex ones requiring basic geometry and calculus training and knowledge, never beyond what is demanded in basic high-school tests.

The questions are of 4 kinds, *scelta multipla*, *completa frase*, *vero/falso* and *numero*:

- *scelta multipla* (multiple choice): the question requires to pick the right answer among four possible ones;
- *vero/falso* (true/false): the question requires to pick the right answer between true and false;
- *numero* (number): the question requires to pick a number that is the correct answer to the question;
- *completa frase* (fill the gap): the question requires to fill one or more missing words to make the text coherent.

Of the four question types, *scelta multipla* and *vero/-falso* are naturally multiple choice, with *scelta multipla* always having 4 possible answers *(A, B, C, D)* and *vero/-falso* always 2, *(true, false)*. Questions of the *numero* type, are not naturally multiple choice, since the answer is a



(a) Invalsi MATE

(b) Invalsi ITA

Figure 2: The distribution of Question types in the Invalsi datasets in (a) for Invalsi MATE and in (b) for Invalsi ITA.

number among all possible ones (some of the answers will be a year, e.g. 1948, while others can be a decimal number of liters of milk, e.g. 0.2), to address this we add 3 extra answers that are realistic but wrong to make the questions multiple choice. Finally, *completa frase* questions, which are only few (20), are too difficult to turn into multiple choice without changing their meaning, and therefore we exclude them.

2.2. Task 2: Language Understanding in Italian (Invalsi ITA)

The second task consists in answering Italian language understanding questions, similarly to task 1, these questions are also appropriate for students between 6 and 18 years old, and they are overall not too difficult to answer. Most of the questions concern a text passage that has to be included in the model context, making this evaluation more costly because the context becomes considerably larger. The text passage is where the difficulty difference between ages is more evident, since it can be a simple and short story for primary school students, while they are generally longer and more involved texts for older students.

The questions are of 4 different types, *scelta multipla*, *binaria*, *domanda aperta* and *altro*:

- *scelta multipla* (multiple choice): the question requires to pick the right answer among four possible ones;
- *binaria* (binary): the question requires to pick the right answer about a binary property of a statement, e.g. True - False, Before - After, etc.
- *domanda aperta* (open question): the question requires to pick the passage in the text that answers the question.
- *altro* (other): A small share of questions belong to open-ended questions with varying scope that are hard to put under a single label.

Similar to Invalsi MATE, this task involves only multiple-choice questions, evaluated through a likelihood approach. Both *scelta multipla* and *binaria* are naturally of this kind, the first with 4 options *(A, B, C D)* and the second with 2 options that change for each question. Both *domanda aperta* and *altro* questions are hard to turn into multiple-choice ones and therefore we discard them. Also for Invalsi ITA, this involves only discarding about 180 questions out of 1297, therefore the task only involves 1117 samples.

3. Data description

3.1. Origin of data

The dataset is built upon the questions from the Invalsi tests of the last 15 years. These tests are administered to students yearly. There are three different Invalsi tests, Language, Mathematics and English. For the scope of this datasets we don't look into the English test, but we limit ourselves to the Italian language and Mathematics ones. The original questions can be accessed here [1]

Some of the questions from the original tests contain visual content as part of the question, we omit these questions since we focus on the language understanding abilities of the models.

[1] https://www.gestinv.it/Index.aspx

Question Type	*ALL*	*scelta multipla*	*vero/falso*	*numero*
N. Questions	400	244	54	102
Model		Accuracy		
mixtral instruct	**0.55**	**0.49**	**0.63**	**0.66**
mistral instruct	0.44	0.34	0.59	0.63
anita 8b dpo	0.47	0.40	0.61	0.55
llama 3 8b instruct	0.48	0.42	0.57	0.58
minerva 3b	0.20	0.22	0.50	0.32
random	0.28	0.25	0.5	0.25

Table 1
Models 0-Shot accuracy on Invalsi MATE, likelihood based evaluation. In **bold** the highest accuracy in each column and underlined the second highest.

Question Type	*ALL*	*scelta multipla*	*binaria*
N. Questions	1117	977	140
Model		Accuracy	
mixtral instruct	**0.80**	**0.82**	**0.69**
mistral instruct	0.49	0.60	0.51
anita 8b dpo	0.71	0.72	0.66
llama 3 8b instruct	0.69	0.70	0.61
minerva 3b	0.30	0.25	0.54
random	0.27	0.25	0.44

Table 2
Models 0-Shot accuracy on Invalsi ITA, likelihood based evaluation. In **bold** the highest accuracy in each column and underlined the second highest.

The data is first collected as is from the webpage and afterwards it is manually checked for errors and inconsistencies from two annotators. The annotators have MSc in Mathematics and Computer Science, which gives them sufficient knowledge to identify issues in the Invalsi MATE questions. For Invalsi ITA, the annotators don't have an appropriate background, however, the questions are simple enough that they can be easily understood and checked for errors by anybody who has completed the mandatory education.

3.2. Data format

The Invalsi MATE dataset has 8 different columns:

- **testo:** this field contains the context needed to answer the questions, it is often empty for Invalsi MATE since most of the context is part of the *domanda* field itself;
- **domanda:** this field contains the question itself, including possible answer options, e.g. for *scelta multipla* questions;
- **risposta:** this field contains the correct answer;
- **test_id:** this field is just an id to identify each sample;
- **tipo:** this field indicates the question type, among *scelta multipla*, *vero/falso*, *numero* and *completa frase*;
- **alt1, alt2 and alt3:** this three fields indicate the alternative values for the *numero* questions since this we chose ourselves and are not indicated in the *domanda* field.

The Invalsi ITA dataset has the same fields as the Invalsi MATE one with the exception that the *testo* field is often present and generally the longest.

We evaluate in a zero-shot fashion just providing the model with question and using a likelihood based method, we pick as the model's answer the one with the highest likelihood among the options available. This is always possible since we have recast all the questions as multiple choice ones. We also don't use chain of thought prompts or similar methods. Since this is the first attempt to build a dataset on mathematical understanding in Italian, currently we evaluate with the simplest approach.

3.3. Detailed data statistics

The data does not have a train and a test split because we have a limited number of samples. The Invalsi MATE split is composed of 420 samples, of which 400 are used in the benchmark, since we exclude the 20 questions marked as *completa frase* since they can´t be made into multiple choice.

Figure 2a shows the percentage of questions of each kind in the dataset, *scelta multipla* has the largest share, 58%, the second most present is *numero*, 24.7% and then *vero/falso* and *completa frase* are fewer. Table 1 has a *random* row that shows the performance if one were to pick random questions, moreover the correct answers for each question type are approximately evenly distributed among labels. Specifically, *scelta multipla* questions have answers distributed as follows, 46 are labelled A, 87 B, 71 C and 40 D, showing a moderate balance. Similarly for *vero/falso* questions there 24 questions with positive answer and 30 with negative answer.

Similarly, Figure 2b shows the percentage of questions of each of the kinds present in Invalsi ITA, *scelta multipla* is by a large margin the most present, composing 76.4% of all the questions, *binaria* is second with 10.9% while *domanda aperta* and *altro* are fewer.

Table 2 shows the performance one can achieve picking answers at random in each split, and moreover the correct answers are evenly distributed for each label also in this dataset. In particular, for Invalsi ITA 254 of the *scelta multipla* questions have answer A, 255 B, 263 C and 205 D which is comparable to the distribution in the Invalsi MATE dataset and similarly does *binaria*.

4. Metrics

Since all the datasets and splits we study have balanced labels, we choose to measure accuracy. In particular, since all the questions in the datasets we propose are multiple choice, it is straightforward to measure accuracy even if there are questions of different kinds, by simply counting $|correct\ answers|/|all\ answers|$.

To see how challenging our benchmark is , we measure four powerful language models based on mistral, mixtral and llama 3, in particular, we measure the performance of *mistral instruct*, *mixtral instruct*, *anita 8b dpo* and *llama 3 8b instruct*.

These models have between 7 and 54 billion parameters, three of them, *mistral instruct*, *anita 8b dpo* and *llama 3 8b instruct* are purely autoregressive transformers, while *mixtral instruct* is a MoE architecture that has 54 billion parameters but only uses 14 billion at inference. We test them all in the same way, using a likelihood based approach.

Table 1 shows the performance of these models on Invalsi MATE on the whole dataset, in the *ALL* column and on each split *scelta multipla*, *vero/falso* and *numero* in the respective columns. *mixtral instruct* is the clear winner among the models we tested, it beats the second best, *llama 3 8b instruct* by 7% accuracy on the entire dataset. On the separate splits, *mixtral instruct* is best overall, however the second best model changes, with *llama 3 8b instruct* being second in *scelta multipla* with a 7% gap, *anita 8b dpo* second on *vero/falso* with a smaller 2% performance gap and *mistral instruct* being second best in *numero* with a 3% gap.

The total accuracy is bound by 55% showing that the Invalsi MATE task is challenging for models of the sizes we tested, up to 54B parameters, and that the performance a model achieves provides valuable insights about how well it can perform mathematical reasoning in Italian.

Table 2 shows the performance of the same models on Invalsi ITA, the model ranking stays the same, with *mixtral instruct* the strongest and *anita 8b dpo* second best. Performance on Invalsi ITA is higher across all fields with *mixtral instruct* achieving 80% accuracy. Unlike what happens for Invalsi MATE, in Invalsi ITA the second best model is the same across the board, *anita 8b dpo* is the second in *scelta multipla* as well as in *binaria* with a performance gap around 10% in all question types.

The accuracy of the best model on all the questions at once is 80% showing that the models we tested perform well in the language understanding in Italian.

5. Conclusions

We propose two Tasks, Invalsi MATE and Invalsi ITA the first for the evaluation of mathematical understanding and the second for the evaluation of language understanding in Italian.

For Invalsi MATE we have collected 420 questions divided into 4 types, *scelta multipla*, *vero/falso*, *numero* and *completa frase* and we evaluate 4 strong language models that are near SOTA in their weight range, *mixtral instruct*, *mistral instruct*, *llama 3 8b instruct* and *anita 8b dpo*. We find that this models are still far from perfect mathematical understanding in Italian with the highest accuracy, achieved by *mixtral instruct* being 55%.

For Invalsi ITA we have collected 1297 questions divided into 4 types, *scelta multipla*, *binaria*, *domanda aperta* and *altro*, we tested the same models also on this benchmark and found that models are stronger at language understanding, with the highest accuracy in this task at 80%, also in this case, achieved by *mixtral instruct*.

Both mathematical and Language understanding are key abilities for LLMs, we believe that our two benchmarks will foster the development of LLMs in Italian and pave the way for new more challenging benchmarks on mathematical and language understanding in Italian.

6. Limitations

The main limitations of the benchmark we propose lies in Task 2, Invalsi ITA we show that the models we test achieve very high accuracy, up to 80% on this benchmark, making it possibly too simple for newer and larger models, nevertheless, current Italian first LLMs are not comparable to larger English-first ones and therefore we believe it can still be valuable in this transitory phase. On the contrary Invalsi MATE is very challenging and it seems that models won't saturate it soon.

We believe that there is a limited risk from contamination from both existing English and Italian tests.

Concerning direct contamination, we were unable to find any web page that would expose the answers openly without needing any sort of authentication, making it difficult to crawl these data automatically, therefore, while the questions might be present in the training set of some of the models, we deem it unlikely that the answers were there too.

Concerning contamination through translation from English, the Invalsi questions are carefully crafted to match the grade of the students that will undertake them, therefore we believe it is unlikely that they are taken from English questions available in other online sources, but rather created specifically for each new annual test.

Acknowledgments

This work was also partially supported by: FAIR (PE00000013) project under the NextGenerationEU programme, partially by the PNRR project ITSERR (CUP

B53C22001770006) and partially by the Project PRIN 2022EPTPJ9 (WEMB – "Word EMBeddings: From Cognitive Linguistics to Language Engineering, and Back"), funded by the Italian Ministry of University and Research (MUR).

References

[1] G. Bolondi, C. Cascella, 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, 2017, p. 14.

[2] A. Costanzo, M. Desimoni, Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using invalsi survey data, Large-scale Assessments in Education (2017). URL: https://doi.org/10.1186/s40536-017-0048-4. doi:10.1186/s40536-017-0048-4.

[3] J. Pietschnig, S. Oberleiter, E. Toffalini, D. Giofrè, 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 (2023) 112345. URL: https://www.sciencedirect.com/science/article/pii/S0191886923002684. doi:https://doi.org/10.1016/j.paid.2023.112345.

[4] A. Esuli, G. Puccetti, The invalsi benchmarks: measuring linguistic and mathematical understanding of large language models in italian, 2024. URL: https://arxiv.org/abs/2403.18697. arXiv:2403.18697.

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

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

[7] H. Liu, Z. Zheng, Y. Qiao, H. Duan, Z. Fei, F. Zhou, W. Zhang, S. Zhang, D. Lin, K. Chen, MathBench: Evaluating the theory and application proficiency of LLMs with a hierarchical mathematics benchmark, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics ACL 2024, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2024, pp. 6884–6915. URL: https://aclanthology.org/2024.findings-acl.411.

[8] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, Training verifiers to solve math word problems, 2021. arXiv:2110.14168.

[9] R. J. Das, S. E. Hristov, H. Li, D. I. Dimitrov, I. Koychev, P. Nakov, Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024. arXiv:2403.10378.

[10] L. Bentivogli, B. Magnini, I. Dagan, H. T. Dang, D. Giampiccolo, The fifth PASCAL recognizing textual entailment challenge, in: Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009, NIST, 2009. URL: https://tac.nist.gov/publications/2009/additional.papers/RTE5_overview.proceedings.pdf.

[11] S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, B. V. Durme, Record: Bridging the gap between human and machine commonsense reading comprehension, 2018. URL: https://arxiv.org/abs/1810.12885. arXiv:1810.12885.

[12] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, SuperGLUE: a stickier benchmark for general-purpose language understanding systems, Curran Associates Inc., Red Hook, NY, USA, 2019.

[13] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. URL: https://aclanthology.org/N18-1101. doi:10.18653/v1/N18-1101.

[14] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. URL: https://aclanthology.org/P18-2124. doi:10.18653/v1/P18-2124.

[15] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: T. Linzen, G. Chrupała, A. Alishahi (Eds.), Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. URL: https://aclanthology.org/W18-5446. doi:10.18653/v1/W18-5446.

[16] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las

Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. arXiv:2401.04088.

[17] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[18] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. E. Tan, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Grattafiori, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Vaughan, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Franco, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Wyatt, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Ozgenel, F. Caggioni, F. Guzmán, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Thattai, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, I. Damlaj, I. Molybog, I. Tufanov, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Prasad, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Huang, K. Chawla, K. Lakhotia, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Tsimpoukelli, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. P. Laptev, N. Dong, N. Zhang, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Li, R. Hogan, R. Battey,

R. Wang, R. Maheswari, R. Howes, R. Rinott, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Kohler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Albiero, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wang, X. Wu, X. Wang, X. Xia, X. Wu, X. Gao, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Hao, Y. Qian, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

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

[20] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.