The Invalsi Benchmarks: measuring the Linguistic and Mathematical understanding of Large Language Models in Italian

Giovanni Puccetti and Maria Cassese and Andrea Esuli
Institute of Science and Technologies of Information "A. Faedo" – CNR Pisa
{giovanni.puccetti,maria.cassese,andrea.esuli}@isti.cnr.it

Abstract

While Italian is a high-resource language, there are few Italian-native benchmarks to evaluate generative Large Language Models (LLMs) in this language. This work presents three new benchmarks: *Invalsi MATE* to evaluate models performance on mathematical understanding in Italian, *Invalsi ITA* to evaluate language understanding in Italian and *Olimpiadi MATE* for more complex mathematical understanding.

The first two benchmarks are based on the Invalsi tests, which are administered to students of age between 6 and 18 within the Italian school system and have been validated by several experts in teaching and pedagogy, the third one comes from the Italian high school math Olympiad.

We evaluate 10 powerful language models on these benchmarks and find that their performance is limited to 71% accuracy on Invalsi MATE, achieved by Llama 3.1 70b instruct and by 88% on Invalsi ITA. For both Invalsi MATE and Invalsi ITA we compare LLMs with the average performance of Italian students to show that Llama 3.1 is the only one which outperforms them on Invalsi MATE while most models do so on Invalsi ITA, we then show that Olimpiadi MATE is more challenging than Invalsi MATE and the highest accuracy, achieved by Llama 3.1 405b instruct accuracy is 45%.

1 Introduction

The evaluation of Large Language Models (LLMs) is a complex task due to the general purpose nature of these systems (Gehrmann et al., 2023). Evaluating different abilities requires both different benchmark datasets and evaluation metrics, there is therefore need for multifaceted evaluation methodologies to perform all-round evaluation of these models.

In this work we propose three benchmarks meant to evaluate language models on mathematical knowledge and language understanding in Italian: *Invalsi MATE*, *Invalsi ITA*, and *Olimpiadi MATE*. The first two are based on the Invalsi tests, public tests that are used to assess students' skills from primary school to high school in the Italian school system, the former is meant to assess a language model's ability to perform math reasoning and the second to assess its language understanding ability. The third one, Olimpiadi MATE, based on the Italian national math Olympiad, is meant to extend Invalsi MATE providing more difficult questions.

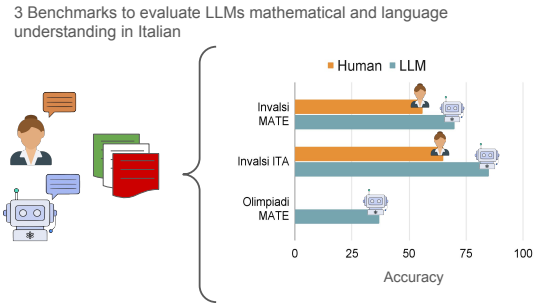


Figure 1: We show that LLMs perform better than human students on Mathematical and Language understanding in Italian.

For questions based on the Invalsi tests, we compare the performance of language models with that of students of different ages across Italy, and we find that language models outperform them in Italian as well, as shown in Figure 1.

The contribution of this work is twofold:

1. **three benchmarks for Math and Language Understanding in Italian** (two for Math, one for Language) that are the first natively Italian benchmarks of this kind; There are benchmarks on translated datasets, however, these are not fully documented or openly available (Jiang et al., 2024).

2. **the evaluation of 10 powerful LLMs on these benchmarks**, including Llama 3.1 405b

Proceedings of the 31st International Conference on Computational Linguistics, pages 6782–6797
January 19–24, 2025.

instruct.

We evaluate 4 kinds of models:

- English pre-trained and English fine-tuned
- English pre-trained and Italian fine-tuned
- Multilingual pre-trained and Multilingual fine-tuned
- Italian pre-trained.

We write English pre-trained to indicate models with pre-training data mostly in English, this does not exclude the presence of non-English data, intentional or not, e.g. Llama includes Wikipedia in many languages (Touvron et al., 2023).

Moreover, we contribute an Italian adaptation of Llama 2 70b, fine-tuned on the Camoscio dataset (Santilli and Rodolà, 2023). This serves as a strong model for the case of models pre-trained in English and fine-tuned in Italian. However, we find that fine-tuning is not as effective as multilingual pre-training when evaluating on non-English languages.

Indeed, those models with mixed language pre-training that on English-based evaluations are comparable to English-first models, perform considerably better on Italian-based evaluations. More generally, we find that larger models trained on more text, e.g. Llama 3.1, outperform multilingual models.

The Invalsi Benchmarks have been added to the CALAMITA suite and the performance of future models will be shown at their webpage.[1] The Invalsi datasets are openly available online.[2]

The rest of the document is organized as follows Section 2 presents the related work and Section 3 describes the benchmarks. Evaluation paradigm along with the models we evaluate are described in Section 4 and the results in Section 4.2, Section 4.3, and Section 4.4. Finally, we compare the models' performance with the results of Italian students in Section 5 and draw conclusions in Section 6.

2 Related work

The development of open, English-first, Large Language Models is improving thanks to recent open weights releases (Touvron et al., 2023; AI@Meta, 2024) and also fully open source ones (Black et al., 2022; Biderman et al., 2023; Groeneveld et al., 2024).

There are also open weights releases of multilingual language models, focused on languages spoken within the European Union (among other continents), French, Spanish, Portuguese, etc., such as Mistral (Jiang et al., 2023) and Mixtral (Jiang et al., 2024), as well as in other non-English languages such as Arabic (Sengupta et al., 2023) and Chinese (Young et al., 2024); we refer to Min et al. (2023) for a more extensive review. There are also models meant to work over hundreds of low-resource languages (Üstün et al., 2024), however, that is beyond the scope of this work.

There are no examples of Language models with 7 billion or more parameters pre-trained in Italian, however initial efforts towards such models are starting, most prominently Minerva, the only previous experiment is IT5 (Sarti and Nissim, 2022) which, while it notably came earlier, it is older and smaller compared to current LLMs.

Nevertheless, several examples of fine-tunes on Italian are available. DanteLLM (Bacciu et al., 2024) is a chat fine-tune version of Mistral and an extension of Fauno (Bacciu et al., 2023). LLamantino is an example of continued pre-training and fine-tuning of LLaMA 2 models on Italian (Basile et al., 2023a) using QLora (Dettmers et al., 2023) the same developers later trained Anita (Polignano et al., 2024) based on Llama 3 (AI@Meta, 2024). Camoscio (Santilli and Rodolà, 2023) is an Italian replica of Alpaca and ExtremITA is a fine-tune dedicated to the 2023 EVALITA challenge (Hromei et al., 2023).

The Occiglot family of models[3] focuses on several languages spoken within Europe, including Italian. Further examples are, Cerbero (Galatolo and Cimino, 2023) and Maestrale .[4] With the exception of Llamantino, which also releases models with 13 and 70 billion parameters, all other fine-tunes are models with at most 7 Billion parameters using quantization (Dettmers et al., 2022) and Lora for fine-tuning (Hu et al., 2022).

2.1 Benchmarks

Along with a few language models pre-trained on Italian, there are also few benchmarks explicitly thought for evaluating LLMs in this language, and in several cases, they are obtained by translating

[1] CALAMITA 2024
[2] huggingface.co/ai4text/Invalsi
[3] https://occiglot.eu/
[4] https://huggingface.co/mii-llm/

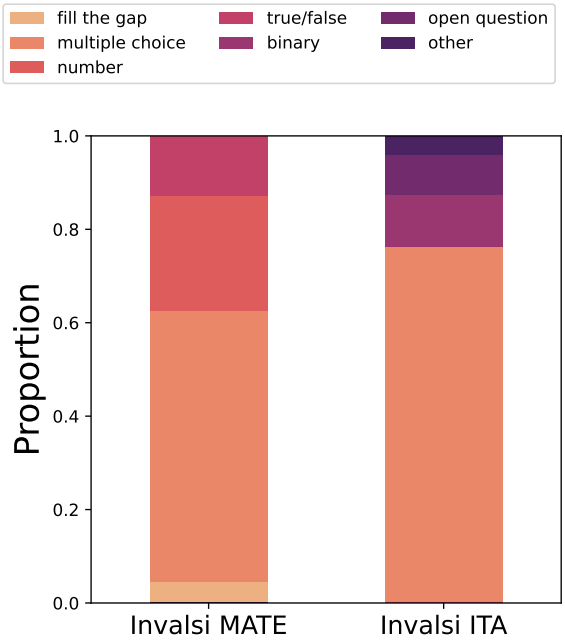


Figure 2: The distribution of Question types in Invalsi MATE and Invalsi ITA.

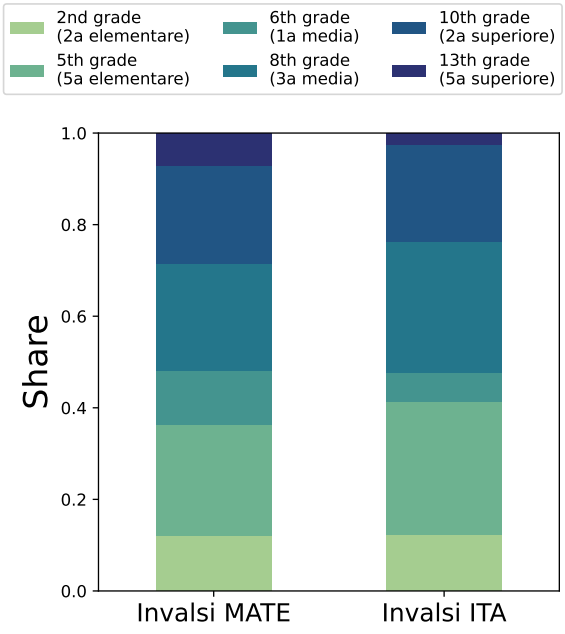


Figure 3: The distribution of questions across school grades in Invalsi MATE and Invalsi ITA. In parenthesis the name of the school grades in Italian.

existing benchmarks for English instead of developing new ones.

To benchmark IT5 Sarti and Nissim (2022) collect a dataset of Italian news for a summarization task, newssum-it, while other benchmarks are obtained by translating existing English datasets, as is done for squad-it (Croce et al., 2018).

The most relevant exception is UINAUIL, a set of benchmarks to evaluate language understanding in Italian (Basile et al., 2023b), based on the tasks presented at several EVALITA campaigns.

There are also a few multilingual benchmarks that include Italian (Hardalov et al., 2020; Das et al., 2024; Zhang et al., 2024), but, to the best of our knowledge, neither mathematical nor linguistic understanding is included. On the contrary, when shifting to English there are many benchmarks focusing on the evaluation of mathematical skills (Cobbe et al., 2021) and scientific knowledge (Dinh et al., 2024).

3 Benchmark Description

The Invalsi tests are country-wide assessments designed to monitor the average performance of students over the years, administered multiple times from primary school through high school.[5] The results of these tests have been used in several population studies (Bolondi and Cascella, 2017; Costanzo and Desimoni, 2017; Pietschnig et al., 2023), however, to the best of our knowledge, their use to benchmark Language Models' performance in Italian is unprecedented. There is work that was preprint after our own (Mercorio et al., 2024), however this work only focuses on linguistic understanding and shows consistent results with our own Invalsi ITA on a smaller dataset.

These tests are of three types: Mathematical Understanding, Language Understanding, and English Understanding. Given our current focus on Italian and the large number of English language understanding benchmarks we discard the last one.

The high school round of Italian Math Olympiad consists of tests designed for students between the ages of 14 and 18. These tests are more challenging than the Invalsi tests and are generally only administered to students willing to take them. The questions are more complex in several ways; they require more reasoning, are often more open-ended, and tend to involve more advanced topics.

To create each benchmark we have collected the data from their original sources and we gathered the full history of questions and answers. Two annotators have manually checked all the samples to control if there were mistakes and if the collected answers were right.

One of the annotators holds an MSc in Mathematics and the other holds an MSc in Computer

[5] https://www.invalsi.it/invalsi/index.php

Question Type		*ALL*	*multiple choice*	*true - false*	*number*
N. Questions		400	244	54	102
Base Model	Model	Accuracy			
llama 3.1 70b	*llama 3.1 70b instruct*	**0.71** (± 0.01)	**0.70** (± 0.01)	**0.64** (± 0.04)	**0.78** (± 0.04)
mistral	*mixtral instruct*	<u>0.55</u> (± 0.02)	<u>0.49</u> (± 0.03)	<u>0.63</u> (± 0.07)	<u>0.66</u> (± 0.04)
	mistral instruct	0.44 (± 0.03)	0.34 (± 0.03)	0.59 (± 0.07)	0.63 (± 0.07)
llama 2 70b	*camoscio 2 70b instruct*	0.49 (± 0.02)	0.43 (± 0.03)	0.59 (± 0.07)	0.62 (± 0.05)
	llamantino 2 70b chat	0.47 (± 0.02)	0.41 (± 0.03)	0.54 (± 0.07)	0.61 (± 0.05)
	llama 2 70b chat	0.43 (± 0.02)	0.40 (± 0.03)	0.48 (± 0.07)	0.52 (± 0.05)
llama 3 8b	*anita 8b dpo*	0.47 (± 0.02)	0.40 (± 0.03)	0.61 (± 0.07)	0.55 (± 0.05)
llama 3.1 8b	*llama 3.1 8b instruct*	0.5 (± 0.01)	0.45 (± 0.01)	0.59 (± 0.04)	0.59 (± 0.07)
minerva	*minerva 3b* 3b	0.20 (± 0.04)	0.22 (± 0.04)	0.50 (± 0.07)	0.32 (± 0.05)
-	*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 <u>underlined</u> the second highest.

science, this gives them sufficient knowledge to evaluate both Invalsi MATE and Olimpiadi MATE while for Invalsi ITA the questions are addressable by any native speaker of Italian that went through the mandatory education.

In all the tests, a small share of the questions also have a visual component. In this work, we focus on textual-only questions and we exclude them. We plan to develop a follow-up benchmark for Italian Visual Language Models (VLMs).

In the future, the annual cycle of Invalsi and Math Olympics allows for a periodic update of the benchmarks based on new test releases. This would provide a continuous stream of test sets resilient to data leakage and allow for an increase in the benchmarks' size.

3.1 Invalsi MATE

The math dataset based on Invalsi is composed of several questions that fall into four caregories:

- *multiple choice*: the student is asked to pick the correct answer among four candidate answers;
- *true - false*: the student is asked to assess whether a given statement is True or False;
- *number*: the student is asked a question that admits a given number as an answer;
- *fill the gap*: the student is asked to fill one or more missing words in a given text, based on logical and mathematical reasoning.

To evaluate LLMs performance on these tasks we use a likelihood-based approach. We compare the likelihood of each possible completion and select the highest one as the answer chosen by the LLM. While only *multiple choice* and *true - false* questions are naturally meant to be evaluated in this way, we recast the *number* questions by manually adding wrong options next to the correct answers.

We exclude the *fill the gap* questions from the evaluation since they are hard to adapt to a multi choice setting. The Invalsi MATE benchmark is composed of a total of 420 questions. Figure 2 shows the share of questions of each type. The *multiple choice* questions make up about 58% of the questions, the second most numerous are *number* questions then *true - false* while there are only 20 *fill the gap* questions.

3.2 Invalsi ITA

Language understanding tests are based on a piece of text that students have to read before answering the questions, this can be as simple as a kid novel for younger students or as complex as an essay or a journal article for older students. After reading the passage, students are presented with a set of questions concerning its content and its linguistic properties.

Similarly to how we organize Invalsi MATE, we group the questions for Italian language understanding in four classes:

- *multiple choice*: the student is asked to pick the correct answer among four candidate an-

Question Type N. Questions		*ALL* 1117	*multiple choice* 977	*binary* 140
Base Model	Model		Accuracy	
llama 3.1 70b	*llama 3.1 70b instruct*	**0.88** (± 0.01)	**0.9** (± 0.01)	**0.75** (± 0.04)
mistral	*mixtral instruct*	<u>0.80</u> (± 0.01)	<u>0.82</u> (± 0.01)	<u>0.69</u> (± 0.04)
	mistral instruct	0.49 (± 0.01)	0.60 (± 0.02)	0.51 (± 0.04)
llama 2 70b	*camoscio 2 70b instruct*	0.78 (± 0.01)	0.78 (± 0.01)	0.67 (± 0.04)
	llamantino 2 70b chat	0.74 (± 0.01)	0.75 (± 0.01)	0.63 (± 0.04)
	llama 2 70b chat	0.72 (± 0.01)	0.73 (± 0.01)	0.64 (± 0.04)
llama 3 8b	*anita 8b dpo*	0.71 (± 0.01)	0.72 (± 0.01)	0.66 (± 0.04)
llama 3.1 8b	*llama 3.1 8b instruct*	0.71 (± 0.01)	0.72 (± 0.01)	0.6 (± 0.04)
minerva	*minerva 3b*	0.30 (± 0.01)	0.25 (± 0.01)	0.54 (± 0.04)
-	*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 <u>underlined</u> the second highest.

swers;

- *binary*: the student is asked to assess a binary property of a statement, e.g. True – False, Before – After, etc.
- *open question*: the student is asked to identify a passage in the text that answers the question;
- *other*: A small share of questions belong to open-ended questions with varying scopes that are hard to put under a single label.

The Invalsi ITA benchmark is composed of a total of 1264 questions. Figure 2 shows the share of questions of each type. The *multiple choice* questions make up more than 76% of the questions, the second in quantity are *binary* questions then *open question*, and finally, there are only 44 *other* questions.

One of the differences between Invalsi MATE and Invalsi ITA is that the questions in the latter set often concern a longer text passage that needs to be processed entirely by the model as context to answer the questions. From a computational perspective, this results in higher memory needs to answer these questions, while from a cognitive perspective, this changes the "abilities" needed to answer them. While Invalsi MATE questions require reasoning skills to be answered, Invalsi ITA questions require the ability to retrieve information from longer texts and leverage general knowledge.

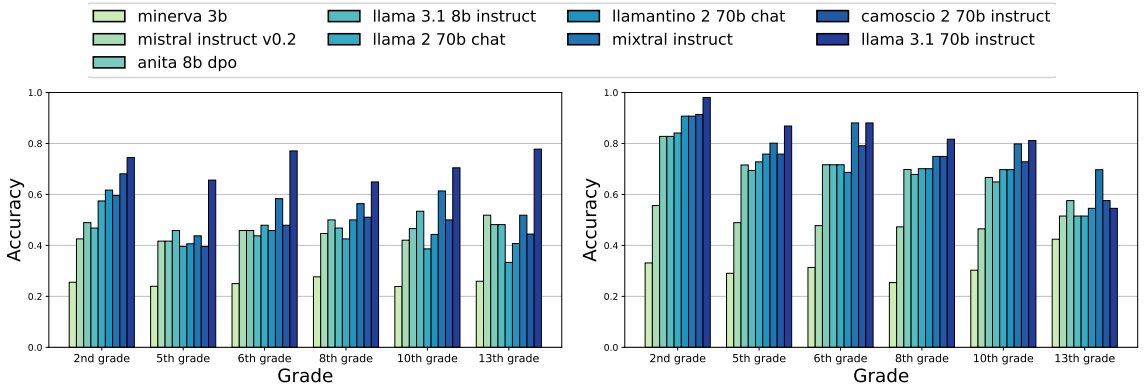
3.3 Olimpiadi MATE

The Olimpiadi MATE benchmark contains only *multiple choice* questions with 5 possible choices A, B, C, D, and E. While this appears to simplify the dataset, these questions are inherently more difficult, as also confirmed in our evaluation. The benchmark consists of 619 questions and as for Invalsi MATE, we discarded those questions involving visual content or with incomplete information.

3.4 Distribution by Grade

The Invalsi tests are taken by students of varying ages between 1st grade, 6 years old, and 13th grade, 18 years old, the distribution of questions by grade is shown in Figure 3. The questions are fairly evenly distributed across all grades in Invalsi MATE while for Invalsi ITA grades 6th and 13th are less present, reflecting the distribution of data in original tests. In Section 4.2 and Section 4.3 we show how models' performance varies across different grades.

The Italian Math Olympiad also has different tests based on the students' age. In particular, 9th and 10th grade students have one set of questions, while 11th, 12th, and 13th grade students have another. We refer to these sets as B (for *Biennio*: first 2 years of high-school in Italy) and T (for *Triennio*: last three years of high-school in Italy) respectively. Set B is composed of 267 questions, and set T is composed of 352 questions. Together they form the dataset used in the Olimpiadi MATE



(a) Invalsi MATE

(b) Invalsi ITA

Figure 4: The performance stratified for different grades, in (a) for Invalsi MATE and in (b) for Invalsi ITA.

benchmark.

4 Evaluation

4.1 Models

We divide the models we evaluate into 4 categories, we add the names we use for each of them in tables in *italics*, see also Appendix A for more details.

English pre-trained and Italian fine-tuned These models are pre-trained on English and then fine-tuned on Italian. We evaluate our own fine-tune of Llama 2 70b on the Camoscio dataset (Santilli and Rodolà, 2023), and we name it Camoscio 2 70b (*camoscio 2 70b instruct*). We also evaluate LLamantino 70B chat (*llamantino 2 70b chat*) which is a fine-tune of LLama 2 70B chat on the ultrachat dataset translated in Italian (Basile et al., 2023a), and the smaller Anita (*anita 8b dpo*) is a model built upon *llama 3 8b instruct*.

Multilingual pre-trained and Multilingual fine-tuned These models are both pre-trained and fine-tuned on multilingual datasets. We consider the two most effective models of this kind that have been trained on Italian and languages close to it. From the Mistral family: Mistral-Instruct-v0.2 (*mistral instruct*) (Jiang et al., 2023) and Mixtral-Instruct-v0.1 (*mixtral instruct*) (Jiang et al., 2024). From the Llama 3.1 family (Touvron et al., 2023): Llama 3.1 8b instruct (*llama 3.1 8b instruct*), Llama 3.1 70b instruct (*llama 3.1 70b instruct*) and Llama 3.1 405b instruct (*llama 3.1 405b instruct*).

English pre-trained and English fine-tuned The developers of these models have intentionally removed all non-English texts in their pre-training corpus, except for selected sources, e.g. Wikipedia. The models of this kind that we check are those from the Llama 2 family, we experiment with Llama 2 70b chat (*llama 2 70b chat*).

Italian pre-trained There is currently only one model pre-trained only on English and Italian, with a focus on Italian, Minerva 3b (*minerva 3b*), however, its largest version encompasses 3 billion parameters and thus it compares poorly to the other larger models we test. We test it nevertheless because in the future, larger models of this same family will be available and it will be a natural step forward to test them on our benchmark.

4.2 Results on Invalsi MATE

Likelihood Based Evaluation We measure models' performance on our dataset using a likelihood-based evaluation method, as done in Gao et al. (2023).

Table 1 shows the results of this evaluation, each column represents the accuracy on a specific type of question, and the *ALL* column the accuracy over the whole set.

The most clear result is how *llama 3.1 70b instruct* outperforms every other model by a large margin, the results in bold in Table 1 show how its accuracy is more than 10% higher than all others across all tasks. However, the Llama 3.1 model family is trained on 15 trillion tokens, reportedly way more than all other models we consider. Looking at the group of the other models trained with comparable training resources, there are two main takeaways:

- **language models pre-trained mostly in English, perform worse than smaller models pre-trained on multilingual data:** *mistral instruct* performs better than comparably large

Dataset Split		*ALL*	B	T
N. Questions		619	267	352
Base Model	Model	Accuracy		
llama 3.1 405b	*llama 3.1 405b instruct*	**0.45** (± 0.02)	**0.46** (± 0.03)	**0.45** (± 0.03)
llama 3.1 70b	*llama 3.1 70b instruct*	<u>0.34</u> (± 0.02)	<u>0.34</u> (± 0.03)	<u>0.34</u> (± 0.03)
mistral	*mixtral instruct*	0.26 (± 0.02)	0.24 (± 0.02)	0.28 (± 0.02)
	mistral instruct	0.27 (± 0.02)	0.29 (± 0.03)	0.26 (± 0.02)
llama 2 70b	*camoscio 2 70b instruct*	0.22 (± 0.02)	0.23 (± 0.03)	0.21 (± 0.02)
	llamantino 2 70b chat	0.25 (± 0.02)	0.23 (± 0.03)	0.26 (± 0.02)
llama 3 8b	*anita 8b dpo*	0.23 (± 0.02)	0.26 (± 0.03)	0.22 (± 0.02)
llama 3.1 8b	*llama 3.1 8b instruct*	0.23 (± 0.02)	0.2 (± 0.02)	0.24 (± 0.02)
	random	0.2	0.2	0.2

Table 3: Models 0-Shot accuracy on Olimpiadi MATE, likelihood based evaluation. In **bold** the highest accuracy in each column and <u>underlined</u> the second highest.

models on Invalsi MATE and so does *mixtral instruct*;

- **further training on Italian does not mitigate this performance gap**: our own *camoscio 2 70b instruct* and *llamantino 2 70b chat* only show marginal gains on mathematical understanding when compared to *llama 2 70b chat* and the same holds for *anita 8b dpo* when compared to *llama 3 8b instruct* showing that fine-tuning on Italian does not appear to help on mathematical understanding.

Indeed, *mixtral instruct* performs better than all other models, including all those based on *llama 2 70b* which are larger in terms of parameters, and all those that had extra training on Italian, *camoscio 2 70b instruct* and *llamantino 2 70b chat*.

Results Break-Down by Grade The Invalsi dataset is stratified by students' grade allowing us to look at model performance at different ages, Figure 4a shows the performance when only answering questions of a given grade. The largest difference between two subsequent grades for a single model is below 10%. However, there is a clear difference between 1st and 2nd grade where models perform better, and the remaining ones: 5th, 6th, 8th, 11th, and 13th.

4.3 Results on Invalsi ITA

We measure the same LLMs also on Invalsi ITA, Table 2 shows their performance on each question type as well as for all of them together.

Similar to what happens for Invalsi MATE *llama 3.1 70b instruct* is the strongest model, however, the performance gap is sensibly lower in this case. While there is a 15% gap between the accuracy of the second best *mixtral instruct* on Invalsi MATE for Invalsi ITA this gap is about 8%. This suggests that **when working on Italian, LLMs can retain reasoning skills acquired in English but suffer a drop in language understanding.**

Excluding *llama 3.1 70b instruct*, similar results also apply to Invalsi MATE: multilingual pre-training provides strong performance improvements in Italian, and the gap is hard to fill with extra training in this language.

Results Break-Down by Grade Invalsi ITA also provides the grade of the students who answered a given question, therefore we can study performance by grade. Figure 4b shows the accuracy of the models on the questions of each grade, unlike for Invalsi MATE we can see a clear descending pattern: LLMs find answering to later grades questions harder and this happens uniformly for all the models we test.

4.4 Results on Olimpiadi MATE

Table 3 shows the evaluation of models on the Olimpiadi MATE benchmark, unlike Invalsi MATE we see that models' accuracy is below 45%, which is only reached by *llama 3.1 405b instruct*, this model is 5 to 50 times larger than all other models we test and costly to use. To run *llama 3.1 405b instruct*, due to its parameter count, we use a

Subject	Question type	Grade					All grades	N. Questions
		2	5	6	8	10		
Mathematics	*multiple choice*	0.55	0.54	0.48	0.58	0.44	0.53	831
	true - false	0.53	0.68	0.51	0.67	0.62	0.65	325
	multi + t - f	0.55	0.58	0.48	0.62	0.51	0.56	1156
Italian	*multiple choice*	0.56	0.63	0.57	0.68	0.63	0.62	1004
	binary	0.61	0.69	0.73	0.74	0.72	0.70	611
	multi + bin	0.57	0.65	0.61	0.70	0.67	0.65	1615

Table 4: Students' accuracy on Invalsi tests, *multiple choice* and and *true - false* questions (together in *multi + t - f*) for Invalsi MATE and *multiple choice* and *binary* questions (together in *multi + binary*) for Invalsi ITA. Source: https://www.gestinv.it/.

large infrastructure, 4 machines each with 4x64GB A100 and therefore we only run it on the Olimpiadi MATE benchmark, the one where other models struggle the most. Looking at the breakdown by B and T, the two sets of questions meant for different age groups, we see that the performance is similar across the two sets. There is a clear best model also for this benchmark *llama 3.1 405b instruct* and *llama 3.1 70b instruct* are the only two models with accuracy above 30%.

Interestingly, *mistral instruct* outperforms *mixtral instruct* on Olimpiadi MATE, although only marginally, unlike on the Invalsi benchmarks. This suggests that *mistral instruct* has undergone deeper fine-tuning for mathematical understanding and confirms that besides conspicuous extra training, multilingual first models perform better in Italian.

We also experiment with different evaluation approaches, either based on pattern matching, see Appendix B or on human evaluation, see Appendix C for more details. Due to space limits, we omit the numerical results from the main body of the work, however, they provide two key insights into the models behaviors:

- Pattern matching-based evaluation is too difficult to assess the performance on our benchmark due to the variability in the answers format, unlike for e.g. GSM8K (Cobbe et al., 2021);
- Human inspection of the results on our benchmarks, although time consuming, shows that models perform comparably when evaluated through likelihood or by assessing the correctness of their textual output.

For examples of the samples in both Invalsi MATE and Invalsi ITA See Appendix D.

5 Comparison with Students' Results

We have put together the aggregated evaluations of students' answers from Invalsi, collected from tens of thousands of students per grade[6], per year (Costanzo and Desimoni, 2017).

We have obtained accuracy scores by question type and grade for both Italian and Mathematics tests, reported in Table 4. It was not possible to exactly select only the questions that compose Invalsi ITA and Invalsi MATE, so this evaluation also accounts for questions that include images.

For this reason, we cannot make a rigorous statistical comparison, yet the values are qualitatively comparable.

Given the caveat, it is interesting to see that on Invalsi MATE only *llama 3.1 70b instruct* performs better than students, with *mixtral instruct* close to them, all other models are worse, most notably on *multiple choice* questions. Differently, on Invalsi ITA most models perform better than students on *multiple choice* questions and instead worse on *binary* ones.

Along grades, the performance of students varies more, with no clear trend, differently from the trend shown by LLMs, more so on Invalsi ITA. This is expected as each grade is a different population and the test for each grade is tailored to that population, while we evaluate each LLM across all grades.

6 Conclusions and Future Work

In this work we introduced three benchmarks, Invalsi MATE, Invalsi ITA and Olimpiadi MATE, to evaluate the performance of LLMs in Italian, on mathematical and language understanding, we collected a total of 1039 questions on mathematical

[6]No data is currently publicly available for the 13th grade.

knowledge and 1249 on language understanding. We measured how well 10 language models perform, including our own fine-tune of LLaMa 2 70B on an Italian dataset. We find that, excluding Llama 3.1 70b (which is reportedly trained on more text than all other models we consider), multilingual models are stronger than English-first ones even when having fewer parameters.

We show that models pre-trained on multiple languages are more accurate in Italian than models pre-trained in English and that fine-tuning on Italian can't fill this performance gap. However, testing *llama 3.1 70b instruct* we show that regardless of the amount of training data, model performance in Italian remains lower than in English. We provide early results showing that LLMs can transfer mathematical reasoning across languages better than language understanding and we find that the gap between *llama 3.1 70b instruct* in Invalsi MATE is larger than in Invalsi ITA. Through evaluations on Olimpiadi MATE we also show that current LLMs are currently not able to consistently solve complex mathematical problems in Italian.

Finally, we report how LLMs perform compared to the population of Italian students to show that they are close to students on Invalsi MATE while they outperform students on Invalsi ITA. Future works will integrate questions about images as well as extensions of the dataset and the validation of future stronger models. Moreover, we intend to continuously update this benchmark along with the yearly Invalsi tests held in schools.

7 Limitations

A limitation of this work is that Invalsi ITA can be almost aced by the most powerful LLMs. Nevertheless, not all the LLMs have top performance on Invalsi ITA and the distribution of performance is indeed varied, making this benchmark appropriate for the early stage development of the Italian LLMs ecosystem, and the exploration of size and resource optimization of state-of-art models. Moreover, since this benchmark is based on real tests, it still provides useful insights into the current performance of LLMs compared to Italian students. Invalsi MATE and specially Olimpiadi MATE show instead to still have margin on the top performing LLMs.

In this work we tested only openly available LLMs. This is motivated by the cost of accessing closed LLMs and also the reduced scientific value of including them into the comparison, due to their lack of specification. To address the potential performance gap with open models, we test *llama 3.1 405b instruct* on the hardest of our benchmarks, whose performance is reportedly comparable to the best closed models [7].

As with most benchmarks designed for LLMs pre-trained on web-scale datasets, there is a potential risk that one or more of the tested LLMs may have seen some of the questions during training. This could lead such LLMs to achieve a spurious better performance. From the descriptions of the training data of the tested LLMs it seems improbable that the content of the tests is explicitly included since they have a very limited diffusion on the web. Moreover, to the best of our knowledge, the answers to the questions of both Invalsi and Math Olympics are accessible only after a registration and a login.

The release of Invalsi MATE, Invalsi ITA, and Olimpiadi MATE exposes them to the risk that future models may be trained explicitly on them. Evaluation of future LLMs should also include a test for the presence of the content of the benchmarks in the training data. On the other hand, the annual nature of these benchmarks allows for continuous updates to the dataset, making it more difficult for models to overfit.

The comparison with students is imperfect since they are evaluated on visual questions too, which are not included in our benchmark, and we can't make a per-question comparison with students, nevertheless, the fact that Invalsi tests are specifically developed to measure the varying of students' performance over the years should mitigate this issue.

Acknowledgements

This work was 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). The authors' opinions do not necessarily reflect those of the funding bodies. We also recognize that the original dataset is developed by Invalsi [8].

[7] https://lmarena.ai/

[8] https://www.invalsi.it/invalsi/index.php

References

AI@Meta. 2024. Llama 3 model card.

Andrea Bacciu, Cesare Campagnano, Giovanni Trappolini, and Fabrizio Silvestri. 2024. DanteLLM: Let's push Italian LLM research forward! In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4343–4355, Torino, Italia. ELRA and ICCL.

Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. 2023. Fauno: The Italian Large Language Model that will leave you senza parole! *arXiv preprint*. ArXiv:2306.14457 [cs].

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023a. Llamantino: Llama 2 models for effective text generation in italian language. *Preprint*, arXiv:2312.09993.

Valerio Basile, Livio Bioglio, Alessio Bosca, Cristina Bosco, and Viviana Patti. 2023b. UINAUIL: A unified benchmark for Italian natural language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 348–356, Toronto, Canada. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 2397–2430, Honolulu, USA. JMLR.org.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Giorgio Bolondi and C 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*, page 14. Franco Angeli, Milano.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Antonella Costanzo 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*.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *Preprint*, arXiv:2403.10378.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. 2024. SciEx: Benchmarking large language models on scientific exams with human expert grading and automatic grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11592–11610, Miami, Florida, USA. Association for Computational Linguistics.

Federico A. Galatolo and Mario G. C. A. Cimino. 2023. Cerbero-7B: A Leap Forward in Language-Specific LLMs Through Enhanced Chat Corpus Generation and Evaluation. *arXiv preprint*. ArXiv:2311.15698 [cs].

Leo Gao, 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. 2023. A framework for few-shot language model evaluation.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Claudiu Daniel Hromei, Danilo Croce, Valerio Basile, and Roberto Basili. 2023. Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy. CEUR.org.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Fabio Mercorio, Mario Mezzanzanica, Daniele Potertì, Antonio Serino, and Andrea Seveso. 2024. Disce aut deficere: Evaluating llms proficiency on the invalsi italian benchmark. *Preprint*, arXiv:2406.17535.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Jakob Pietschnig, 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.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: An Italian Instruction-tuned LLaMA. *Preprint*, arxiv:2307.16456.

Gabriele Sarti and Malvina Nissim. 2022. IT5: Large-scale text-to-text pretraining for italian language understanding and generation. *ArXiv preprint 2203.03759*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024. M3exam: a multilingual, multimodal, multilevel benchmark for examining large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Identifier	Huggingface Name	N. Params	(pre-trainig) Language	(fine-tuning) Language
mixtral instruct	mistralai/Mixtral-8x7B-Instruct-v0.1	47 B	multilingual	multilingual
mistral instruct	mistralai/Mistral-7B-Instruct-v0.2	7 B	multilingual	multilingual
llama 2 7b chat	meta-llama/Llama-2-7b-chat-hf	7 B	English	English
llama 2 70b chat	meta-llama/Llama-2-70b-chat-hf	70 B	English	English
llamantino 2 70b chat	swap-uniba/LLaMAntino-2-70b-hf-UltraChat-ITA	70 B	English	Italian
camoscio 2 70b instruct	ai4text/camoscio-70-b	70 B	English	Italian
llama 3 8b instruct	meta-llama/Meta-Llama-3-8B-Instruct	8 B	English	English
llama 3 70b instruct	meta-llama/Meta-Llama-3-70B-Instruct	70 B	English	English
llama 3.1 8b instruct	meta-llama/Meta-Llama-3.1-8B-Instruct	8 B	English	English
llama 3.1 70b instruct	meta-llama/Meta-Llama-3.1-70B-Instruct	70 B	English	English
llama 3.1 405b instruct	meta-llama/Meta-Llama-3.1-405B-Instruct	70 B	English	English
anita 8b dpo	swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	8 B	English	Italian
minerva 3b	sapienzanlp/Minerva-3B-base-v1.0	3 B	Italian	N/A

Table 5: Model naming table.

A Model Naming Summary

In Table 5 we report the names and identifiers along with their Huggingface model name, number of parameters and the most occuring language in pre-training and fine-tuning data.

B Pattern Matching Evaluation

Table 7 shows the models accuracy on the Invalsi MATE benchmark, both the global accuracy when measuring all the question types together, as well as the accuracy on each separate question type. We notice how performance is heavily dependent on the type of questions, indeed the models perform very poorly on "Fill the gap" tasks, as shown by the "completa frase" columns and on True - False tasks they are also mostly close to random performance when not worse. On the contrary, they show strong performance on both multiple choice and number answers where the strongest model, *mixtral*, achieves up to 61.76% accuracy. This evaluation is done by extracting the answers automatically (i.e., matching answer-template patterns with the text).

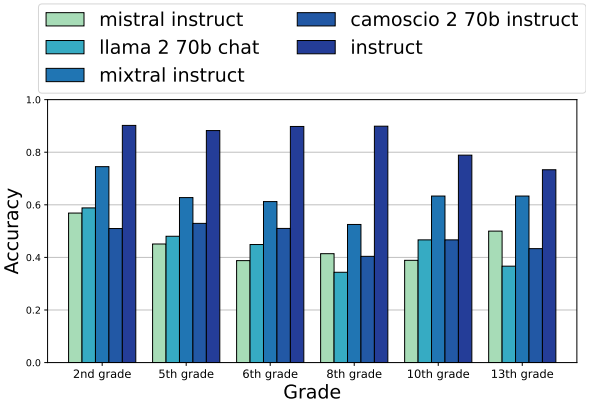


Figure 5: Performance of different Language Models on Invalsi MATE per grade level humanly assessed.

However, seeing the weak performance on True -

Question Type	*ALL*	*fill the gap*	*multiple choice*	*number*	*true - false*
N. Questions	420	20	244	54	102
Model	Accuracy				
llama 3 70b instruct	0.86	0.63	0.86	0.83	0.98
mixtral instruct	0.62	0.37	0.63	0.63	0.65
camoscio 2 70b instruct	0.48	0.32	0.47	0.49	0.52
llama 2 70b chat	0.45	0.37	0.43	0.44	0.56
mistral instruct v0.2	0.44	0.21	0.44	0.42	0.56

Table 6: Models 0-Shot accuracy on Invalsi MATE, human evaluation.

model	All	sentence completion	multiple choice	number	true/false
mixtral instruct	0.53	0.00	0.60	0.42	0.57
llama 2 70b chat	0.42	0.00	0.46	0.38	0.48
mistral instruct v0.2	0.42	0.00	0.46	0.35	0.52
llama 2 13b chat	0.24	0.00	0.27	0.12	0.46
llama 2 7b chat	0.22	0.00	0.27	0.20	0.48
camoscio 2 70b instruct	0.13	0.00	0.14	0.96	0.39

Table 7: Models 0-Shot accuracy on Invalsi MATE, pattern matching-based evaluation.

False questions and after a manual inspection of the sentences we performed the more accurate human-made evaluation reported in Appendix C.

C Invalsi Mate Human Evaluation

Following likelihood based evaluation, we are interested in knowing the ability of Language models in answering questions by properly generating a complete answer, this is mostly interesting for Invalsi MATE since these questions require explicit reasoning and moreover, the evaluation on Invalsi ITA is too time consuming due to the long text passages providing the contexts to the questions, therefore we limit this analysis to Invalsi MATE.

For this open-ended setting we evaluate a subset of the language models tested so far: *mistral instruct*, *mixtral instruct*, *llama 2 70b chat*, *camoscio 2 70b instruct*, and *llama 3 70b instruct*, we devise a distinct prompt for each question type and prompt all the tested language models to assess how well they understand Mathematical reasoning and Language in Italian.

We perform the evaluation in 0-shot fashion, i.e. each model is asked to answer the question and we use the chain of thought approach (Wei et al., 2022) in that we prefix the language models answer by adding the words "Ragioniamo passo passo" (literally: "Let's reason step by step") to have it output explicit reasoning steps before providing an answer.

We remark that we don't use a chat based approach where we inject further requests one after the other, because some of the models we test are not meant for chat but are only instruction fine-tuned.

Table 6 shows the accuracy achieved by LLMs we manually inspected. Models' ranking is kept and we observe that models fine-tuned in Italian have a lesser ability to clearly frame the answer, i.e. "the correct answer is A" (in Italian, "la risposta esatta è A"), with respect to those fine-tuned in English or multilingual datasets, even though the correct answer is returned.

We argue that the main reason for this issue is the content and more importantly the size of the Italian fine-tuning datasets, which are smaller and of lower quality.

We also look at the performance distribution by grade, reported in Figure 5 which shows similar trends to what observed for likelihood based evaluations

D Questions Examples

Table 8 and Table 9 report examples of the questions in Invalsi MATE and Invalsi ITA respectively.

Multiple Choice	*True - False*
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	**Testo** Se moltiplichi per 2 un numero naturale e dal risultato sottrai 1, ottieni sempre un numero pari. **Domanda** Vero o Falso?
Translation	Translation
Text Elisa left her house in the morning at 8:15 am. Elisa came back in the afternoon at 1:15 pm **Question** How long was Elisa outside? A. 5 hours; B. 7 hours; C. 9 hours	**Text** If you multiply a natural number by 2 and subtract 1 you always get an even number. **Question** True or False?
Number	***Fill the gap***
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?	**Testo** Luca lancia due dadi a sei facce non truccati. **Domanda** Completa la frase inserendo una delle seguenti espressioni. - maggiore della - minore della - uguale alla La probabilità che la somma dei punti sia 12 è .... probabilità che la somma sia 2.
Translation	Translation
Text Filippo says: to find out the number of my t-shirt add one tens and six units to 4 **Question** What is the number on Filippo's t-shirt?	**Text** Luca tosses two fair six-sided dice. **Question** Finish the sentence adding one of the following options - higher than - smaller than - equal to "The probability that the sum of the two dice is 12 is ... probability that the sum is 2."

Table 8: **Examples of Invalsi MATE samples.**

Multiple Choice	*Binary*
Testo Il titolo dice che Polipetto ha un problema e l'inizio del racconto spiega di che cosa si tratta. **Domanda** Qual è il problema di Polipetto? A. Non gli piace più la sua casa B. Non può più entrare in casa sua C. La sua casa si è riempita di animali D. La sua casa non è più ordinata come prima	**Testo** Grazie all'incontro con il grande Oceano qualcosa cambia in Polipetto. **Domanda** Com'è Polipetto PRIMA di parlare con Oceano e DOPO avere parlato con lui? Metti una crocetta per ogni riga. A. Polipetto si fa coraggio: Prima/Dopo B. Polipetto è confuso: Prima/Dopo C. Polipetto pensa che gli altri siano più bravi di lui: Prima/Dopo D. Polipetto si fida della sua idea: Prima/Dopo
Translation	Translation
Text The title says that Polipetto has a problem and the beginning of the story exaplains what it is. **Domanda** Which one is Polippeto's problem? A. He doesn't like his house B. He can't enter his house C. His house is filled with animals D. His houes is not tidy anymore	**Text** After talking with great Oceano something changes in Polipetto. **Question** How is Polipetto BEFORE talking to Oceano and AFTER speaking with him? Pick one for each line. A. Polipetto motivates himself: BEFORE/AFTER B. Polipetto is confused: BEFORE/AFTER C. Polipetto believes that others are better than he is: BEFORE/AFTER D. Polipetto trusts his idea: BEFORE/AFTER
Open Question	***Other***
Testo Che cosa ci dice il racconto a proposito dell'escursione? Completa la sintesi che segue, inserendo la parola appropriata in ogni spazio. **Domanda** "Durante l'escursione alcuni ragazzi non si accorgono di quanto camminano perché cantano e scherzano. All'arrivo, dopo esattamente cinque ore di strada. ... (1) si siede perché è affaticato e i ... (2) gli fanno male; posa vicino a sé lo zaino con la ... (3) dentro"	**Testo** "Doveva assolutamente parlarne con un amico". Nel testo non c'è scritto che cosa ha detto Polipetto a questo amico, ma dal racconto si può capire. **Domanda** Che cosa può avergli detto Polipetto?
Translation	Translation
Text What does the story say about the fieldtrip? Complete the summary that follows, adding the appropriate word in every gap. **Question** "During the fieldtrip some kids loose track of how long they walked because they sing and joke. When they arrive, after walking for 5 hours, ...(1) sits down because he is tired and his ...(2) hurt; he lays his backpack next to himself with the ...(3) inside. Miss Salici finds the lake gorgeous, while the ...(4) are not impressed."	**Text** "He had to talk about it with a friend". The story does not mentioned what Polipetto said to his friend, but it can be understood from the story. **Question** What could have Polipetto said to him?

Table 9: Examples of Invalsi ITA samples.