A Benchmark to Evaluate LLMs' Proficiency on Italian Student Competencies

Fabio Mercorio[1,2], Mario Mezzanzanica[1,2], Daniele Potertì[3], Antonio Serino[3], and Andrea Seveso[1,2] (✉)

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca {name.surname}@unimib.it
[2] CRISP Research Centre, University of Milano-Bicocca
[3] Department of Economics, Management and Statistics, University of Milano-Bicocca

Abstract. Recent advancements in Large Language Models (LLMs) have significantly enhanced their ability to generate and manipulate human language, highlighting their potential across various applications. Evaluating LLMs in languages other than English is crucial for ensuring their linguistic versatility, cultural relevance, and applicability in diverse global contexts, thus broadening their usability and effectiveness. We tackle this challenge by introducing a structured benchmark using the INVALSI tests, a set of well-established assessments designed to measure educational competencies across Italy. Our study makes three primary contributions: First, we adapt the INVALSI tests as a benchmark for automated LLM evaluation, rigorously adapting the test format to suit automated processing while retaining the essence of the original tests. Second, we provide a detailed assessment of current LLMs, offering a crucial reference point for the academic community. Finally, we visually compare the performance of these models against human results. Additionally, our benchmark is publicly available and provided with a comprehensive evaluation suite[4], ensuring that the benchmark remains a current and valuable resource relevant for advancing industrial-strength NLP applications.

Keywords: Large Language Models · Benchmark · Evaluation

1 Introduction

In recent years, Large Language Models (LLMs) have emerged as a pivotal advancement in the field of Natural Language Processing (NLP) and Artificial Intelligence (AI) [9]. Model evaluation is paramount but difficult since there are various important qualities to consider: models should be precise, resilient, fair, and efficient, among others [26]. Developing language models that function effectively across diverse global languages and evaluating them remains a significant and ongoing challenge [41]. The currently available models often perform highly in English but are lacking in underrepresented languages [38]. This is due to factors such as the scarce and lower quality available data [21], smaller contributing communities, and Anglo-centric cultural bias in development [42]. In the current landscape, there is a pressing need for a reliable tool to evaluate models' proficiency in the Italian language, particularly to assess their ability to align with the cultural and linguistic nuances critical for effective deployment in industrial contexts.

The INVALSI (National Institute for the Evaluation of the Education and Training System) test has been crucial in Italy's educational assessment since the 2005-2006 academic year. It evaluates students' competencies in subjects like the Italian language and mathematics at various educational stages. The primary goal is to assess linguistic proficiency, focusing on reading comprehension, grammatical knowledge, and lexical competence [45].

INVALSI tests use real-world language tasks to measure understanding of texts, appropriate vocabulary use, and application of grammatical rules [13,43,17]. The design ensures progressive complexity suitable for each educational level, providing fair and challenging assessments. These tests offer transparent benchmarks for student performance, guiding instructional strategies [31].

Since the test covers a wide range of linguistic and comprehension skills [14], using it to evaluate LLMs can provide a detailed view of a model's proficiency in handling real-world, nuanced language tasks designed for human learners. The test's structured and standardised nature makes it an excellent benchmark for comparing different LLMs with questions culturally and contextually relevant to Italian speakers. However, our findings are applicable across all languages since they assess various general capabilities, such as word formation and text comprehension abilities. Additionally, since it is designed for multiple educational stages, it offers a range of complexity and challenges. This aspect can gauge an LLM's capability at various difficulty levels, reflecting its potential scalability and adaptability across simpler to more complex linguistic tasks.

Given these robust evaluation criteria, this paper aims to establish a benchmark for assessing the performance of large language models by leveraging the INVALSI framework.

1.1 Contributions

The contributions of this work are fourfold:

[4] https://github.com/Crisp-Unimib/INVALSI-Eval-Suite

1. We structure the INVALSI test, a notable national test for Italian students, as a publicly available evaluation benchmark for LLMs[5], making it a valuable resource for those aiming to integrate Italian language services powered by LLMs into their business operations;
2. We perform an in-depth analysis of existing models, using our benchmark, establishing a reference for the research community;
3. We visually display results across several important metrics and compare models' performances to human standards, pinpointing the strengths and weaknesses;
4. We make the dataset available along with an evaluation suite, ensuring the replicability of our results and allowing anyone to test their model on the benchmark[6].

The remainder of the paper is structured as follows: Section 2 presents related work in the state of the art; Section 3 details our data curation process for creating the benchmark; Section 4 displays the results of multiple models tested against this benchmark. Section 5 discusses these results and identifies limitations; Section 6 concludes the paper and outlines proposals for future work.

2 Related Work

A Large Language Model is a deep learning model trained on vast amounts of text data to develop a sophisticated understanding of language structures and semantics. Leveraging the transformer architecture [46], LLMs employ self-attention mechanisms to process sequential data efficiently, exemplified by models such as GPT [35].

Multilingual models. LLMs have shown multilingual capabilities based on their training on multilingual data [44,24] and vocabulary [32,10,25]. GPT-3 and its successors have shown different capabilities in several languages [3] since their training corpora are, in part, composed of non-English texts. Most recently, smaller size models [20,44], due to the inclusion of multilingual data in the training process, have shown emerging capabilities in German, French, Spanish and Italian but not performing as well in the most prominent training language. Given the peculiarity of the Italian language and the lack of consistency of multilingual models in the Italian language, [15] pioneered the first Italian-adapted GPT-2-based model. The development of low-resource adaptation techniques [18,16] enabled the adaptation of larger models to Italian. Recently [39] instruction-tuned LLaMa with the Alpaca dataset translated into Italian, while [4] implemented Parameter Efficient Fine Tuning (PEFT) [28] using synthetically generated, machine-translated data. Additionally, [5] applied PEFT to LLaMa2 across multiple scales (7B, 13B, 70B). Most recently, [33] adapted an 8B LLaMa3 model to Italian via PEFT.

Available benchmarks. Available benchmarks aim to evaluate commonsense reasoning [11,47], multi-step mathematical reasoning [12], Question-Answering [27] and reading comprehension capabilities [36]. The Italian NLP community lacks the depth of original language evaluation benchmarks compared to the English community. Some natively English benchmarks, such as [47,11], are commonly used to evaluate LLMs in Italian after being automatically translated. Benchmarks natively Italian are less common. [6] propose a Unified Benchmark for Italian Natural Language Understanding that covers textual entailment, Event detection and classification, factuality classification, sentiment polarity classification, irony detection and hate speech detection. [22] proposes a collaborative benchmark on 13 tasks. Both benchmarks focus on classification-based tasks and do not explore LLM properties, such as common-sense reasoning. Another Italian benchmark is represented by [23], which concentrates solely on Italian news text summarisation abilities. Additionally, [29] introduced a specialised benchmark for evaluating LLMs on Italian driving license knowledge, demonstrating domain-specific evaluation approaches. More recently, [40] proposed a culture-aware benchmark specifically designed to assess LLMs' understanding of Italian cultural contexts and nuances. In a previous paper [34], the authors structured the INVALSI data to create a benchmark; however, unlike our work, they did not include any open-ended questions. These benchmarks lack a wide range of possible scenarios to evaluate LLMs, thus not allowing a comprehensive evaluation [26].

3 INVALSI Benchmark Curation

We have collected from public sources 58 unique tests, divided into 141 unique units, with 2114 questions and 2808 unique items. Some questions are subdivided into multiple items, each requiring a specific answer.

Data for this study was sourced from the Gestinv[7] database [7]. This database, widely used in Italian educational research and teacher development programs, includes questions from national assessments since 2008, as well as related test materials, statistical reports, and educational tools to enhance the understanding of student learning outcomes across Italy.

[5] We use a subset of tests, handpicked from different years and educational levels, ensuring that we exclude those with questions that are difficult to rephrase or that require analysing images.

[6] `https://github.com/Crisp-Unimib/INVALSI-Eval-Suite`

[7] `https://www.gestinv.it/Index.aspx`

The questions' formatting is sometimes not adequately structured for LLM evaluation; for instance, it is occasionally impossible to automatically transcribe the questions into structured fields, necessitating further inspection of images and PDFs. For this reason, we also collected corresponding PDF files and images. Manual inspection was required to ensure accuracy. In cases where questions involved graphical elements, we modified them into a multiple-choice format that was more suitable for analysis. For example, if the task required a student to find and underscore a word, we reformulated the question to allow selection from multiple choices. Similarly, if the task involved drawing a line between two groups of concepts—a common task for younger students—we rephrased it to include choosing the correct association from given options. Generally, we aimed to adapt the questions to a format that allows the model to select the correct answer from a pool of choices if it aligns with the original question type. Fig. 1 shows a few illustrative question examples.

Illustrative Examples

Multiple Choice (MC) Question

Question:
In the sentence: "Livia was running in the park when a strong storm broke out," what are the events indicated by the two verbs?

Options:

- A. They are contemporary and have the same duration
- B. They are contemporary and indicate habitual actions
- C. The first event occurs during the second event
- D. The second event occurs during the first event (✓)

Multiple Complex Choice (MCC) Question

Question:
Read this sentence: "The night bird made such an acute sound that frightened the inhabitants of the forest very much."

Indicate whether *The* is a noun or not.

Options:

- A. It's a noun
- B. It's not a noun (✓)

Unique Response (RU) Question

Question:
Where would you put the letter h? Indicate whether or not it is necessary instead of **.

**avevo perso l'autobus così arrivai tardi a scuola.

Options:

- A. Necessary
- B. Not necessary (✓)

Fig. 1: Illustrative examples of the INVALSI benchmark for each question format. (✓) indicates the correct answer. **Note:** the original questions are in Italian, and the translation in English is purely for illustrative purposes.

3.1 Dataset Characteristics

We have selected 11 tests comprising 31 unique units, 405 questions, and 618 items from the above data. A test consists of two or more different units; each question can have more than one item to answer. The sample of tests was chosen through manual inspection, aiming to include a variety of grades and years, and avoiding those with questions that require image inspection or contain questions that would be difficult to reformulate for language model comprehension.

Tab. 2 shows the macro area distribution in our benchmark.

Table 1: Distribution of tests, questions, and items by educational grade and question format.

Test Distribution by Grade			
School Grade	# Tests	# Questions	# Items
2nd (Primary School)	2	34	72
5th (Primary School)	2	73	115
6th (Middle School)	2	87	118
8th (Middle School)	2	86	88
10th (High School)	2	75	134
13th (High School)	1	50	91
Question Distribution by Format			
Format		# Questions	# Items
MC		337 (83.2%)	340
MCC		35 (8.6%)	228
RU		33 (8.1%)	50

Table 2: Distribution of questions by section and macro area.

Section	Macro Area	# Questions
Text comprehension	Reconstruct the meaning of the text, locally or globally	177 (43.7%)
	Locate and identify information within the text	108 (26.7%)
	Reflect on the content or form of the text, locally or globally, and evaluate them	33 (8.1%)
Reflection on the language	Lexicon and semantics	29 (7.2%)
	Morphology	24 (5.9%)
	Syntax	18 (4.4%)
	Word formation	7 (1.7%)
	Textuality and pragmatics	5 (1.2%)
	Spelling	4 (1.0%)

"Locate and identify information within the text" is used for all questions aiming to evaluate the capability of identifying various types of information within the provided context. *"Reconstruct the meaning of the text, locally or globally"* assesses how well one can infer and reconstruct the text's context and the encyclopedic knowledge it conveys. Lastly, *"Reflect on the content or form of the text, locally or globally, and evaluate them"* questions aim to evaluate the ability to interpret texts and their shape, expressing an evaluation. The remaining macro areas are designed and structured to evaluate grammatical knowledge. *"Word formation"* aims to assess knowledge about base words and their derivatives, and *"Lexicon and semantics"* aims to assess knowledge about the semantic relationship between words. The questions belonging to *"Morphology"* category aim to check the competencies with several lexical categories (noun, adjective, etc.) and sub-categories (possessive adjective, proper name, etc.). In contrast, using accents and apostrophes, upper and lowercase letters, etc., is evaluated by the questions categorised as *Spelling*. All the questions within *Syntax* aim to assess the correctness of syntactic rules of Italian written language, and *"Textuality and pragmatics"* aims to evaluate signs of text organisation and cohesion phenomena.

The questions come in three distinct formats, which are:

- *Multiple Choice (MC)*: composed of a question with several answer options, among which only one is correct. It is the most common question format in the selected tests, comprising 337 questions (83.2% of the total) and 340 items. Some questions require the selection of two distinct options, both of which must be correct. The answer choices are typically four, labelled A, B, C, and D.
- *Multiple Complex Choice (MCC)*: composed of input questions and multiple items to answer. It is the second most common type of question, with 38 (9.4%) instances and 228 items. For each item, one answer from among the two or more available options must be selected, and only one is correct. The question is deemed correct only if all the items are rightly answered.
- *Unique Response (RU)*: involves open-ended questions in which there are no options or suggestions and where only one answer is considered correct (with sometimes a limited number of possible variants). We found 33 (8.1%) RU questions and 50 items in the selected tests.

3.2 Evaluation

The diversity in question formats ensures a comprehensive evaluation of the models' capabilities. 82.5% of the questions follow a standard multiple-choice format with closed options, 9.4% of the questions are multiple binary choices, requiring

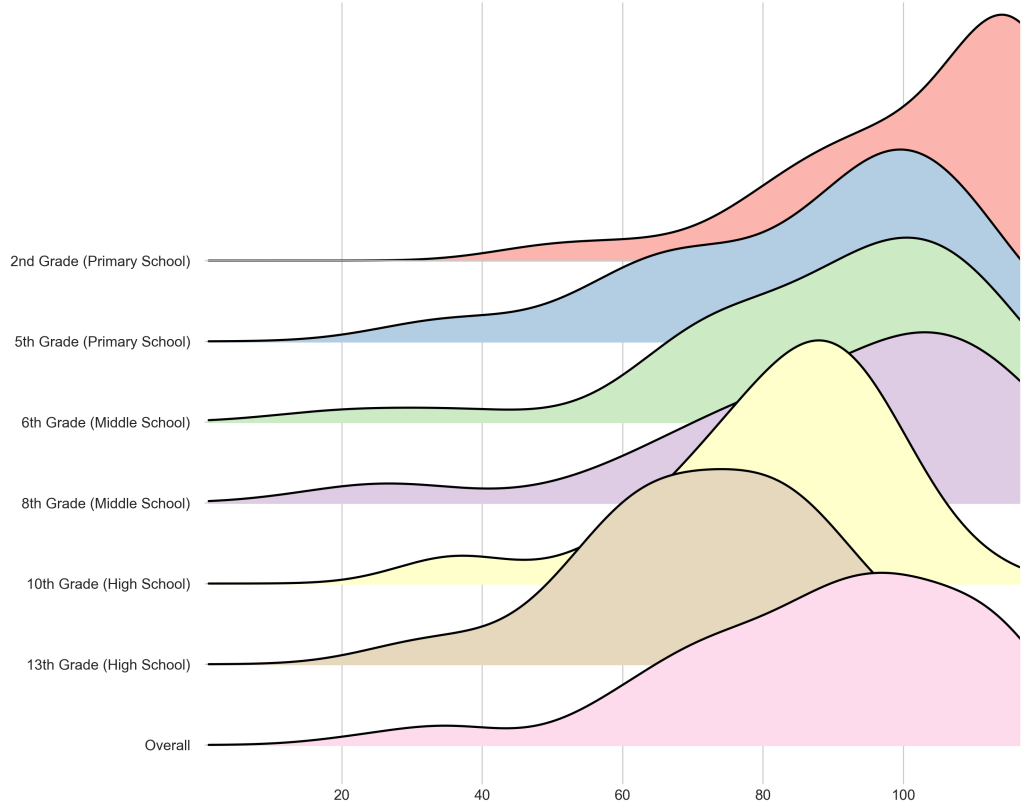

Fig. 2: Visualising the accuracy of various models across different school grades. Each layer represents a different grade level, from 2nd grade in primary school to 13th grade in high school, showing the distribution of performance accuracy for each grade.

the model to affirm or deny statements, and 1% of the questions require specific handling, such as adding punctuation to one particular sentence. As seen in Tab. 1, 91.9% of the questions in the benchmark dataset created are MC and MCC. Consequently, evaluation involves verifying whether the generated answer includes the target answer, which is accomplished through the use of regular expressions. In the case of questions with multiple items, such as MCC, the question is only considered correct if all the items are answered correctly. Our approach is consistent with standard LLM evaluation frameworks, such as OpenAI's simple-evals[8]. We generate responses for open-source models and API-based LLMs and then compare these to the labelled correct answers. We use a temperature of 0 to ensure deterministic output in this setting. Each evaluation has been run exactly once. The benchmark tests models in a zero-shot scenario, requiring them to understand and correctly answer based on their general knowledge or the provided narrative.

Specific strategies for RU questions were developed to cater to their unique requirements. These questions require more than simply selecting an available option; therefore, the prompts were tailored with additional or specific instructions to ensure accurate responses. Each function for generating prompts checks for the presence of context and choices, incorporating them into the prompt if available. This provides the LLM with all pertinent information to answer the question accurately.

Specific evaluation methods include *word matching*, extracting words or phrases and comparing them directly to the answers, making it especially effective for questions that require specific terms or phrases. Another strategy is *pattern matching*: employing regular expressions to detect patterns in the model's output, aiding in evaluating syntactic or grammatical responses. Lastly, we use *BERTScore* [48] to assess the semantic content of answers with a threshold of 0.7 for correctness. This method ensures that responses capture the intended meaning, not just the exact wording, which is vital for complex language tasks where paraphrasing or diverse expressions may still be correct. This threshold has been empirically validated across all model answers. BERTScore utilises BERT's contextual embeddings to accurately evaluate the semantic similarity between responses and reference texts.

4 Results

Model selection criteria. We evaluate a variety of notable foundational and fine-tuned models, chosen based on the following characteristics: (i) *Parameter threshold.* Models with at least three billion parameters are included to ensure substantial complexity and language comprehension capacity. (ii) *Temporal range.* Focuses on models published from

[8] https://github.com/openai/simple-evals

2023 onwards to capture recent advancements and influential models. (iii) *Institutional source.* Considers models from prominent organisations like OpenAI and Meta. (iv) *Popular Italian models.* Includes models specifically trained or fine–tuned in Italian.

For closed-source models, we include OpenAI's GPT-4o and GPT-4o-mini [1], both recognised for their advanced language capabilities. Additionally, we consider Anthropic's Claude series, which includes Haiku and Sonnet, each excelling in text generation tasks. Also part of our evaluation is Google's Gemini Pro 1.5 and 2.0 flash [37], as well as Gemma 3 27B. Our selection of open-source models includes Mistral 7B [19] and Mixtral [20]. Furthermore, we examine Meta's LLaMA 3 series models in three different sizes [2] (405B, 70B, 8B). For models specifically tuned to the Italian language, we include Minerva 7B [30], a foundational model trained from scratch in Italian, as well as Almawave's Velvet 14B[9]. We also consider LLaMAntino 3 [33], a model fine-tuned from LLaMa 3, popular models Llama-3.1-8b-Ita[10] and maestrale-chat-v0.4[11]. We did not inspect older versions of Italian models due to their lower performance. We evaluated the open-source models in bf16 format on an NVIDIA H100 80GB PCIe GPU, using VLLM with its default OpenAI-compatible server.

We also categorise the models into three categories by their size. **Small (S)** models have fewer than 8 billion parameters, or for the closed models accessible via API; they cost less than 0.50$ per million input tokens. **Medium (M)** models can go up to 70 billion parameters for open-source versions and cost less than 5$ per million input tokens for proprietary APIs. **Large (L)** models exceed these limits.

4.1 Model Performance

Given the various dimensions available, we present an overall accuracy distribution for each model to conduct our evaluation. The school grade is the most critical variable influencing our analysis; in Fig. 2, we provide a plot illustrating how accuracy distributions vary across different grades. This visual does not detail specific numbers but instead offers a general sense of how performance shifts with grade level, with a more detailed analysis to follow. Another intriguing dimension to consider is the impact of model size on performance. In Fig. 3, we present the distribution of scores segmented by the model size.

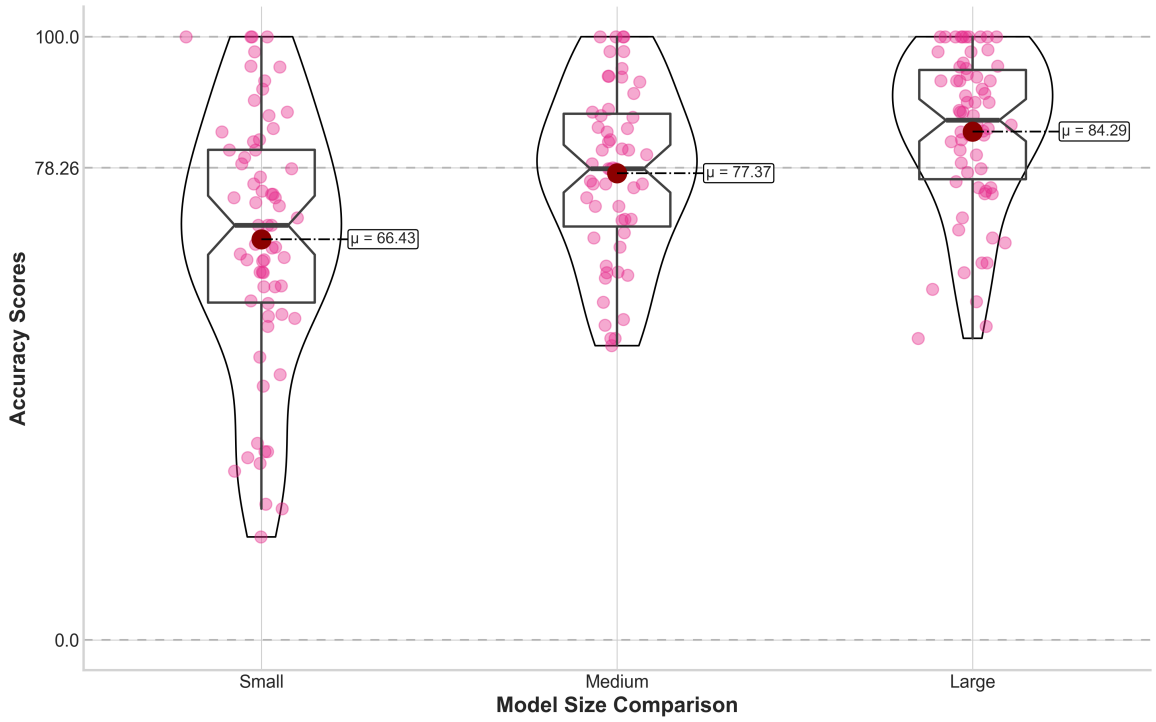

Fig. 3: Distribution of accuracy scores of language models categorised by size: small, medium, and large. Each plot represents the distribution of accuracy scores within each category, with individual data points highlighted, each representing a test taken by a model, and the mean accuracy marked by a horizontal line.

Detailed analysis of question format and macro areas. We then delve into a detailed analysis of how question format influences performance. In Tab. 3 and 4, we present the accuracy scores for each model, stratified by both school grade

[9] Almawave/Velvet-14B
[10] DeepMount00/Llama-3.1-8b-Ita
[11] mii-llm/maestrale-chat-v0.4-beta

and question format. Due to the stratification and the limited number of questions in some categories, extreme values such as 100 or 0 are more attainable in the sections with few items. The number of questions for each category is indicated in the table headers. The model average is in the last column of Tab. 4. Similarly, Tab. 5 shows the performance comparison of AI models across linguistic macro areas.

Table 3: Performance (accuracy %) comparison of AI models across school grades and question formats for grades 2 to 6.

School Grade	2nd Grade (Primary School)		5th Grade (Primary School)			6th Grade (Middle School)		
Question Format (#)	MC (32)	MCC (2)	MC (61)	MCC (6)	RU (6)	MC (72)	MCC (6)	RU (9)
claude-3.7-sonnet	96.3	100.0	100.0	89.8	66.7	64.3	95.2	50.0
claude-3.5-haiku	88.9	100.0	75.0	85.7	58.3	57.1	76.2	0.0
gpt-4o	92.6	100.0	75.0	95.9	75.0	71.4	81.0	37.5
gpt-4o-mini	81.5	0.0	100.0	81.6	58.3	57.1	66.7	0.0
gemini-pro-1.5	92.6	100.0	100.0	81.6	58.3	64.3	83.3	25.0
gemini-2.0-flash	92.6	0.0	100.0	83.7	66.7	71.4	83.3	12.5
gemma-3-27b-it	88.9	100.0	75.0	83.7	33.3	57.1	69.0	12.5
Mistral-Large	88.9	0.0	100.0	85.7	66.7	64.3	81.0	12.5
mistral-nemo	77.8	0.0	75.0	73.5	41.7	71.4	73.8	0.0
llama-3.1-405b	93.8	100.0	100.0	93.9	66.7	71.4	83.3	37.5
llama-3.3-70b	88.9	100.0	75.0	85.7	58.3	78.6	81.0	0.0
llama-3.1-8b	64.2	0.0	50.0	67.3	25.0	64.3	61.9	0.0
Velvet-14B	60.5	0.0	50.0	71.4	25.0	35.7	57.1	12.5
Llama-3.1-8b-ITA	66.7	0.0	75.0	77.5	33.3	57.1	64.3	0.0
LLaMAntino-3-8B	63.0	0.0	50.0	67.3	25.0	42.9	52.4	0.0
maestrale-chat-v0.4	67.9	100.0	50.0	71.4	33.3	50.0	61.9	0.0
Minerva-7B	23.5	0.0	0.0	40.8	16.7	7.1	33.3	0.0
Models Avg	78.1	47.1	73.5	78.6	47.5	58.0	70.9	11.8

Table 4: Performance (accuracy %) comparison of AI models (cont.), for grades 8 to 13 and overall average.

School Grade	8th Grade (Middle School)			10th Grade (High School)			13th Grade (High School)		All Grades
Question Format (#)	MC (81)	MCC (1)	RU (4)	MC (49)	MCC (12)	RU (14)	MC (42)	MCC (8)	Overall
claude-3.7-sonnet	96.3	100.0	100.0	89.8	66.7	64.3	95.2	50.0	92.3
claude-3.5-haiku	88.9	100.0	75.0	85.7	58.3	57.1	76.2	0.0	81.2
gpt-4o	92.6	100.0	75.0	95.9	75.0	71.4	81.0	37.5	90.1
gpt-4o-mini	81.5	0.0	100.0	81.6	58.3	57.1	66.7	0.0	78.3
gemini-pro-1.5	92.6	100.0	100.0	81.6	58.3	64.3	83.3	25.0	85.9
gemini-2.0-flash	92.6	0.0	100.0	83.7	66.7	71.4	83.3	12.5	87.7
gemma-3-27b-it	88.9	100.0	75.0	83.7	33.3	57.1	69.0	12.5	78.3
Mistral-Large	88.9	0.0	100.0	85.7	66.7	64.3	81.0	12.5	85.2
mistral-nemo	77.8	0.0	75.0	73.5	41.7	71.4	73.8	0.0	73.6
llama-3.1-405b	93.8	100.0	100.0	93.9	66.7	71.4	83.3	37.5	90.4
llama-3.3-70b	88.9	100.0	75.0	85.7	58.3	78.6	81.0	0.0	83.7
llama-3.1-8b	64.2	0.0	50.0	67.3	25.0	64.3	61.9	0.0	61.5
Velvet-14B	60.5	0.0	50.0	71.4	25.0	35.7	57.1	12.5	57.5
Llama-3.1-8b-ITA	66.7	0.0	75.0	77.5	33.3	57.1	64.3	0.0	66.2
LLaMAntino-3-8B	63.0	0.0	50.0	67.3	25.0	42.9	52.4	0.0	58.3
maestrale-chat-v0.4	67.9	100.0	50.0	71.4	33.3	50.0	61.9	0.0	64.7
Minerva-7B	23.5	0.0	0.0	40.8	16.7	7.1	33.3	0.0	28.6
Models Avg	78.1	47.1	73.5	78.6	47.5	58.0	70.9	11.8	64.9

4.2 Comparison with Human Respondents

In evaluating the performance of language models, a critical comparison arises between the responses generated by these models and those of human respondents. We aim to provide insights into the capabilities of language models relative to average human performance.

Not every test we had included the percentage of human accuracies. Specifically, data was available from one test for grade 2; for grades 5 and 6, there were accuracies from two tests each; and for grades 8 and 10, accuracies were available from one test each. Unfortunately, no data on human accuracies was available for grade 13.

In Fig. 4, we compare human and model performances. The red lines represent the median of human answers, set at 59.8, to delineate which classes of models perform above this benchmark. This division creates four quadrants: both perform well, neither perform well, humans perform better, and models perform better.

Table 5: Performance (accuracy %) comparison of AI models across macro areas. Categories are abbreviated as: *LI*: Locate and identify information within the text. *RM*: Reconstruct the meaning of the text, locally or globally. *RC*: Reflect on the content or form of the text, locally or globally, and evaluate them. *WF*: Word formation. *LS*: Lexicon and semantics. *MO*: Morphology. *SP*: Spelling. *SY*: Syntax. *TP*: Textuality and pragmatics.

Section	Text Comprehension			Reflection on the Language						Both
Macro Area (#)	LI (108)	RM (177)	RC (33)	WF (7)	LS (29)	MO (24)	SP (4)	SY (19)	TP (5)	Overall
claude-3.7-sonnet	94.4	92.7	81.8	100.0	96.6	87.5	50.0	100.0	100.0	92.3
claude-3.5-haiku	82.4	85.9	81.8	57.1	69.0	83.3	0.0	66.7	100.0	81.2
gpt-4o	89.8	93.8	84.8	100.0	82.8	91.7	0.0	88.9	100.0	90.1
gpt-4o-mini	78.7	84.2	84.8	71.4	62.1	66.7	0.0	61.1	100.0	78.3
gemini-2.0-flash	89.8	89.8	78.8	100.0	79.3	87.5	0.0	94.4	100.0	87.7
gemini-pro-1.5	91.7	88.1	81.8	71.4	82.8	66.7	25.0	83.3	100.0	85.9
gemma-3-27b-it	82.4	81.4	72.7	42.9	62.1	79.2	0.0	83.3	100.0	78.3
Mistral-Large	88.0	87.0	81.8	85.7	79.3	79.2	25.0	88.9	80.0	85.2
mistral-nemo	79.6	81.9	78.8	57.1	48.3	41.7	0.0	44.4	100.0	73.6
llama-3.1-405b	89.8	93.2	87.9	85.7	86.2	83.3	50.0	94.4	100.0	90.4
llama-3.3-70b	88.0	89.3	84.8	71.4	72.4	70.8	0.0	66.7	60.0	83.7
llama-3.1-8b	63.9	72.3	66.7	14.3	41.4	25.0	0.0	38.9	80.0	61.5
Velvet-14B	63.9	63.8	66.7	0.0	34.5	33.3	0.0	38.9	80.0	57.5
Llama-3.1-8b-ITA	71.3	73.4	75.8	28.6	55.2	33.3	0.0	38.9	60.0	66.2
LLaMAntino-3-8B	61.1	66.7	69.7	0.0	41.4	29.2	0.0	44.4	40.0	58.3
maestrale-chat-v0.4	66.7	71.2	66.7	57.1	44.8	54.2	0.0	38.9	100.0	64.7
Minerva-7B	33.3	33.3	21.2	0.0	13.8	12.5	0.0	33.3	20.0	28.6
Models Avg	77.3	79.3	74.5	55.5	61.9	60.3	8.8	65.0	83.5	64.9

5 Discussion

Model performance across stratifications. In analysing the performance results of these models, it is evident that models with a higher number of parameters generally demonstrate superior performance compared to those with fewer parameters, as illustrated in Figure 3. The figure reveals that smaller models exhibit greater variance and dispersion in their accuracy scores than medium and large models. They achieve an average accuracy score of 66.43%. Medium-sized models have a slightly lower average accuracy, compared with the average total accuracy (78.26%), achieving an average score of 77.37%. While the larger sized models score well above average, achieving an accuracy of 84.29%.

Examination of Tables 3 and 4 reveals significant variability in model performance across various school grades.

Models tend to perform better in lower grades, while showing lower accuracy in higher ones. In particular, grade 6 and grade 13 are the grades in which the models have the most difficulty in answering the present questions correctly.

A comparative analysis of LLMs across various macro areas is shown in Tab. 5. A notable strength of these models is their ability to reconstruct the meaning of text, locally and globally (RM), with an overall accuracy of 79.3%. Conversely, none of the models consistently performed well in the spelling (SP) category, with an overall accuracy of 8.8%. A surprising result is the 83.5% accuracy achieved by the models in question in the Textuality and Pragmatics (TP) category. An example of this category is shown in Fig. 5.

Overall, LLMs exhibit superior performance on average in Text Comprehension tasks compared to Reflection on the Language tasks. This aligns with previous findings in the field of Language Understanding, where these language models can excel at understanding context and drawing inferences based on large contexts because of their generative pre-training and discriminative fine-tuning [35]. Conversely, syntax and morphology tasks require precise, rule-based understanding and application. Although models can produce grammatically correct text, they often encounter difficulties with tasks that demand explicit knowledge of linguistic rules and higher levels of reasoning. None of the models tested could correctly answer all or even the majority of the SP category questions. These questions presented a classical Italian writing task:

The letter "h" should be placed correctly to answer this question and form the appropriate Italian words. The letter "h" is essential in Italian for distinguishing between certain homophones. The correct placement would differentiate "ho" (I have) and "ha" (he/she has), but in this context, the correct form is "a scuola" (at school), meaning no "h" is needed. Therefore, the sentence reads: "avevo perso l'autobus così arrivai tardi a scuola" (I missed the bus so I arrived late at school). We observed that some of the larger closed models correctly answered one or two of the four spelling questions.

Impact of model size on performance. Large closed-source models achieve superior accuracy on benchmarks, successfully addressing approximately 80-85% of the tasks. The Claude class model exhibits the highest accuracy with *Claude Sonnet 3.7* at 92.3%, while OpenAI's *GPT 4o* model shows an accuracy of 90.1%. In the domain of open weights models, *llama-3.1-405b* achieves an accuracy of 90.4%, while *Mistral-Large* produce a 85.2% accuracy score. Among the Italian pre-trained models, *Minerva-7B* shows an accuracy rate of 28.6% and *Velvet-14B* shows an accuracy of 57.5%.

Whereas among the improved Italian models, *Llama-3.1-8b-ITA*, which is an improved variant of LaMA-3-8b, demonstrates improved performance, achieving a 66.2% success rate, an increase of 4.7 percentage points over the base model.

Overall, the inference for this benchmark with closed-weight models incurred a cost slightly below $50. The dataset comprises approximately 620,000 input million tokens, yielding an average output of 17,000 tokens per model. As of the

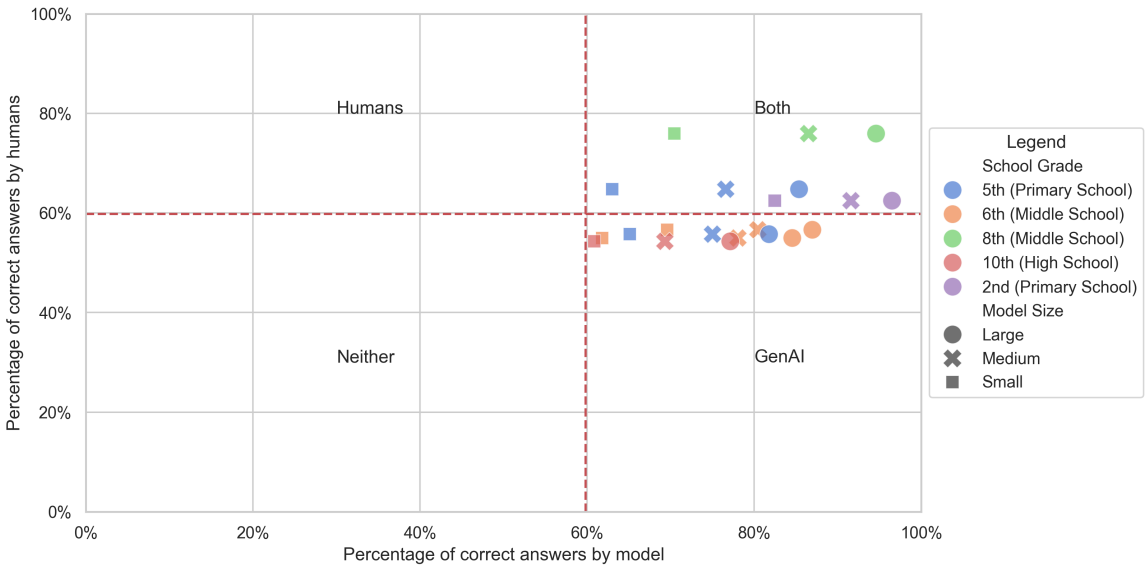

Fig. 4: Scatter plot visualising the accuracy of both human respondents and language models on various tests across different grade levels. The red lines represent the median accuracy of human answers at 59.8%. The graph is divided into four quadrants to categorise the performance: top-right quadrant ("Both"), where both humans and models perform well; top-left quadrant ("Humans"), where humans outperform models; bottom-right quadrant ("GenAI") where models outperform humans; and bottom-left quadrant ("Neither") where neither models nor humans perform well. Each symbol represents the average performance for each model size on a test, and colour coding corresponds to the educational grade level, providing an overview of where AI competes or lags behind human performance. Multiple data points with the same colour and symbol are shown wherever multiple tests for the same school grade exist.

Textuality and Pragmatics Question

In the following sentence, insert the missing punctuation marks. Rewrite the sentence with the missing marks.

Mother called Little Red Riding Hood and told her please go to grandma; bring her these things the butter the eggs and the sugar

Fig. 5: Example of a Textuality and Pragmatics question.

inference date, the costliest model was ChatGPT 4-o, priced at $5 per million input tokens and $15 per million output tokens.

Comparison of models with human respondents. When evaluating language models, comparing their responses to human respondents is crucial. However, human accuracy data was not uniformly available: it was present for one test in grade 2, two tests each in grades 5 and 6, and one test each in grades 8 and 10, with no data for grade 13. In Fig. 4, we visualise those comparisons. The red lines, representing the median human score 59.8%, identify which models perform above or below this level. This creates four quadrants: both perform well, neither perform well, humans perform better, and models perform better. It is interesting to notice that, while we assume that LLMs' performance would vary linearly with task difficulty, human cognitive development does not follow a linear path in individuals but occurs in stages marked by times of discontinuity [8], especially during adolescence.

Spelling Question

Where would you place the letter *h*?
If needed, write it in the box.

□avevo perso l'autobus così arrivai tardi □a scuola.

Fig. 6: Example of a Spelling question.

5.1 Limitations

Data availability. The dataset obtained from Gestinv includes all the INVALSI tests on the Italian language; however, a few questions (3 or 4) were missing from certain tests. Moreover, in some tests, particularly those labelled as simulations, a few questions were missing multiple-choice options, rendering the questions unclear. Some metadata was wrongly labelled. These minor issues were identified and rectified through manual intervention.

Potential shortcomings in complex answer evaluation. A subset of questions (seven in total) posed a significant challenge in the evaluation due to the requirement for subjective judgment to determine the correctness of the answers. These questions necessitate that the generated answers be semantically relevant to the target answers provided as references. The complexity arises because semantic relevance is not always easily quantifiable, leading to potential inconsistencies in assessment. To address this, we employed BERTscore [48] to establish an empirical threshold where answers with a BERT score greater than 0.70 were considered correct, while those below this threshold were deemed incorrect. While this method provided a systematic evaluation approach, it has limitations. In practice, this method has been manually validated to work well in all present cases, and future cases will be carefully monitored.

6 Conclusion and Future Work

This research paper introduces a new benchmark for evaluating large language models by structuring the Italian INVALSI tests. Key contributions include establishing a structured benchmark for the Italian language, extensively assessing current LLMs, and comparing model performances across various dimensions. Due to the increased complexity of language and cognitive tasks at higher levels, models perform better on tasks for lower school grades than on higher ones. Models excel in text comprehension but find reflecting on the Italian language harder. Larger models outperform smaller ones, even those pre-trained and fine-tuned for the Italian language, indicating that extensive training data and complex architectures help handle language task nuances better. We also release the *data* and *evaluation suite* to allow anyone to test their model on our benchmark at `https://github.com/Crisp-Unimib/INVALSI-Eval-Suite`.

Looking ahead, the research aims to expand the benchmark's scope and utility by *(i) incorporating mathematics and multimodal capabilities* to test the models' abilities to handle linguistic, quantitative, and visual information. *(ii) Increasing the test size* to enhance the robustness of evaluations, reduce variance, and provide a more comprehensive assessment of LLMs' linguistic capabilities.

A Ethical Considerations

The dataset is composed entirely of publicly available test questions and does not include any confidential information, personal data, or non-public communications. All data and supplementary materials used in the collection process are free from personally identifiable information or sensitive content. An ethical review process was unnecessary since the dataset is derived solely from public tests and does not involve human subjects or private data. However, potential misuse risks exist, such as using benchmark results to support or oppose the development of native LLMs specifically tailored to the Italian language. Careful consideration is advised to prevent misinterpretations or unintended consequences when applying the evaluation outcomes.

B Resource Availability Statement

Our benchmark is accessible at `https://doi.org/10.5281/zenodo.15553471` under the MIT license, with no IP-based or other restrictions. Additionally, the code used for the evaluation is available on GitHub[12].

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AI@Meta: Llama 3 model card (2024), `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`
3. Armengol-Estapé, J., Bonet, O.d.G., Melero, M.: On the multilingual capabilities of very large-scale english language models. arXiv preprint arXiv:2108.13349 (2021)
4. Bacciu, A., Trappolini, G., Santilli, A., Rodolà, E., Silvestri, F.: Fauno: The italian large language model that will leave you senza parole! arXiv preprint arXiv:2306.14457 (2023)
5. Basile, P., Musacchio, E., Polignano, M., Siciliani, L., Fiameni, G., Semeraro, G.: Llamantino: Llama 2 models for effective text generation in italian language. arXiv preprint arXiv:2312.09993 (2023)

[12] `https://github.com/Crisp-Unimib/INVALSI-Eval-Suite`

6. Basile, V., Bioglio, L., Bosca, A., Bosco, C., Patti, V.: 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). pp. 348–356 (2023)
7. Bolondi, G., Gambini, A., Ferretti, F.: Il database gestinv delle prove standardizzate invalsi: Uno strumento per la ricerca: Alcuni esempi di utilizzo nell'ambito della matematica. In: I dati INVALSI: Uno strumento per la ricerca, pp. 43–48. Franco Angeli (2017)
8. Carey, S., Markman, E.M.: Cognitive development. In: Cognitive science, pp. 201–254. Elsevier (1999)
9. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X.: A survey on evaluation of large language models (2023), http://arxiv.org/abs/2307.03109
10. Chung, H.W., Garrette, D., Tan, K.C., Riesa, J.: Improving multilingual models with language-clustered vocabularies. arXiv preprint arXiv:2010.12777 (2020)
11. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457 (2018)
12. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems, 2021. URL https://arxiv. org/abs/2110.14168 (2021)
13. Corsini, C.: La validità di contenuto delle prove invalsi di comprensione della lettura. Italian Journal of Educational Research (10), 46–61 (2013)
14. Corsini, C., Losito, B.: Le rilevazioni invalsi: a che cosa servono? Cadmo: giornale italiano di pedagogia sperimentale: 2, 2013 pp. 55–76 (2013)
15. De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., Guerini, M.: Geppetto carves italian into a language model. arXiv preprint arXiv:2004.14253 (2020)
16. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems **36** (2024)
17. Guzzo, G.: La competenza grammaticale nelle prove invalsi (2023)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
19. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
20. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
21. Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., et al.: Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics **10**, 50–72 (2022)
22. Lai, M., Menini, S., Polignano, M., Russo, V., Sprugnoli, R., Venturi, G., et al.: Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian. In: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy (2023)
23. Landro, N., Gallo, I., La Grassa, R., Federici, E.: Two new datasets for italian-language abstractive text summarization. Information **13**(5), 228 (2022)
24. Li, J., Zhou, H., Huang, S., Cheng, S., Chen, J.: Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. Transactions of the Association for Computational Linguistics **12**, 576–592 (2024)
25. Liang, D., Gonen, H., Mao, Y., Hou, R., Goyal, N., Ghazvininejad, M., Zettlemoyer, L., Khabsa, M.: Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. arXiv preprint arXiv:2301.10472 (2023)
26. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
27. Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958 (2021)
28. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., Bossan, B.: Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft (2022)
29. Mercorio, F., Potertì, D., Serino, A., Seveso, A., et al.: Beep-best driver's license performer: A calamita challenge. In: CEUR WORKSHOP PROCEEDINGS. vol. 3878 (2024)
30. Orlando, R., Moroni, L., Cabot, P.L.H., Barba, E., Conia, S., Orlandini, S., Fiameni, G., Navigli, R., et al.: Minerva llms: The first family of large language models trained from scratch on italian data. In: Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024) (2024)
31. Pastore, S., Freddano, M., et al.: "questione di feedback": dati invalsi e pratiche di valutazione in classe. In: I dati INVALSI: uno strumento per la ricerca, pp. 89–100. FrancoAngeli (2017)
32. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502 (2019)
33. Polignano, M., Basile, P., Semeraro, G.: Advanced natural-based interaction for the italian language: Llamantino-3-anita (2024)
34. Puccetti, G., Cassese, M., Esuli, A.: The invalsi benchmarks: measuring the linguistic and mathematical understanding of large language models in Italian. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp. 6782–6797. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), https://aclanthology.org/2025.coling-main.453/
35. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
36. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
37. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)

38. Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., et al.: Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2021)
39. Santilli, A., Rodolà, E.: Camoscio: An italian instruction-tuned llama. arXiv preprint arXiv:2307.16456 (2023)
40. Seveso, A., Potertì, D., Federici, E., Mezzanzanica, M., Mercorio, F., et al.: Italic: An italian culture-aware natural language benchmark. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), April 29-May 4, 2025. vol. 1, pp. 1469–1478 (2025)
41. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)
42. Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., Radev, D., et al.: You reap what you sow: On the challenges of bias evaluation under multilingual settings. In: Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models. pp. 26–41 (2022)
43. Tóth, Z.: Riflettere sulle parole: la formazione delle parole nelle prove invalsi. Lingue antiche e moderne **12**, 277–298 (2023)
44. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
45. Trinchero, R.: Il servizio nazionale di valutazione e le prove invalsi. stato dell'arte e proposte per una valutazione come agente di cambiamento. Form@ re-Open Journal per la formazione in rete **14**(4), 34–49 (2014)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
47. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830 (2019)
48. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)