BAMBI Goes to School: Evaluating Italian BabyLMs with Invalsi-ITA

Luca Capone[1,*,†], Alice Suozzi[2,†], Gianluca E. Lebani[2,3,†] and Alessandro Lenci[1,†]

[1]CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Via Santa Maria 36, 56126 Pisa, Italy
[2]QuaCLing Lab, Dipartimento di Studi Linguistici e Culturali Comparati, Università Ca' Foscari Venezia, Dorsoduro 1075, 30123 Venice, Italy
[3]European Centre for Living Technology (ECLT), Ca' Bottacin, Dorsoduro 3911, 30123 Venice, Italy

Abstract

This paper explores the impact of ecologically and cognitively plausible data on the training of language models. It builds on prior work [1, 2] integrating child-directed speech, curriculum learning and instruction tuning to train Italian BabyLMs. To evaluate our BabyLMs, we compare their performance (trained on fewer than 100M words using various techniques) with that of native Italian Large Language Models using the Invalsi-ITA [3] benchmark, designed to evaluate Italian students on text comprehension and linguistic abilities. The goal is to assess whether cognitively motivated training approaches (Curriculum Learning based on Child-Directed speech and child-friendly data), which are crucial for meaningful comparison between human learners and computational systems [4], yield greater efficiency than standard methods.

Keywords

Italian BabyLM, Invalsi-ITA benchmark, LM Evaluation, Text Comprehension, Italian Grammar

1. Introduction

Even though Language Models (LMs) have taken research in linguistics and cognitive science by storm, their meaningful application in these fields still faces significant challenges. In order for LMs to be useful and informative for understanding language and cognition, several plausibility criteria must be met [5, 6, 7]. Among them, the most important are the amount of input received during training and the number of trainable parameters. A growing body of empirical evidence shows that beyond a certain model size and amount of training data, the probability distributions generated by LMs diverge from human-like patterns and become poor predictors of psycholinguistic measures, such as eye-tracking data [8, 9]. In contrast, smaller models trained on a limited amount of data appear to align more closely with human reading strategies. This observation is consistent with findings from the BabyLM Challenge, which demonstrate that models trained on child-directed speech and capped at 100 million words can achieve strong syntactic competence [10, 11]. In addition to model size and training data volume, other plausibility criteria should be considered. These include the quality of the input (such as child-directed speech) and the manner in which it is presented, for instance through Curriculum Learning (CL). Moreover, the standard language modeling objective differs substantially from the discursive and interactive exchanges children engage in with adults and peers [4]. In short, approximating child language learning conditions requires attention to multiple dimensions.

This study aims at investigating the impact of such dimensions on LMs' development of linguistic skills. Specifically, we examine the effectiveness of training Italian BabyLMs using child-directed speech, curriculum learning, and instruction tuning—techniques inspired by human language acquisition to the purpose of assessing whether these cognitively grounded methods lead to improved performance compared to conventional training approaches, particularly when working with limited data. To this end, we evaluate our BabyLMs against native Italian Large Language Models using the Invalsi-ITA benchmark, which is focused on text comprehension and linguistic knowledge.

The paper is structured as follows: first, an overview of related works is provided in Section 2. Section 3 is dedicated to the description of the models' evaluation. The models are presented in Section 3.1, whilst in Sections 3.2 and 3.3 the Invalsi-ITA benchmark, used for the evaluation, and the procedure followed to assess the models' abilities are described. The results of the evaluation are detailed in Section 3.4 and discussed in Section 3.5. Finally, some conclusions are drawn in Section 4.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.

† For the specific purposes of Italian Academy, Luca Capone is responsible for Sections 3.1, 3.3 and 3.5, Alice Suozzi for Section 2, 3.2, and 3.4 Alessandro Lenci for Section 1 and Gianluca E. Lebani for Section 4.

luca.capone@fileli.unipi.it (L. Capone); alice.suozzi@unive.it (A. Suozzi); gianluca.lebani@unive.it (G. E. Lebani); alessandro.lenci@unipi.it (A. Lenci)

0000-0002-1872-6956 (L. Capone); 0000-0002-5215-7742 (A. Suozzi); 0000-0002-3588-1077 (G. E. Lebani); 0000-0001-5790-4308 (A. Lenci)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-11-30

2. Related Works

Two lines of research are particularly relevant to our goals, as they represent two sides of the same coin: the first focuses on the quality and quantity of training data necessary for BabyLMs to develop linguistic abilities; the second concerns the evaluation of BabyLMs through the creation or adaptation of benchmarks originally designed to assess the linguistic competence of human speakers.

Regarding the first aspect, several studies have explored training models on datasets that are comparable—both in size and in linguistic nature—to the input typically received by children during early development (e.g., [12, 13, 14]). These works show that while a large volume of data is essential for achieving strong performance on standard Natural Language Understanding tasks, a significantly smaller amount is sufficient for acquiring core syntactic knowledge. In addition to data quantity and quality, the importance of curriculum learning strategies and model architecture optimization has also been highlighted [10].

On the evaluation front, several benchmarks have been developed over the years (e.g., [15, 16, 17]). While these benchmarks are effective tools for comparing models against each other, they are not well-suited for comparing models to human language abilities, especially those of children. Although some studies have directly addressed this gap (e.g., [18]), they have not yet produced large-scale, standardized benchmarks for this purpose.

For the Italian language, to the best of our knowledge, only two benchmarks currently enable both model-to-model and model-to-human comparisons. The first is BaBIEs [1], a benchmark derived from the adaptation of four standardized tests originally designed to assess the semantic and syntactic competence of Italian-speaking children. The second is Invalsi-ITA [3, 19], described in Section 3.2, which aims to evaluate text comprehension and linguistic abilities in Italian students from primary through high school.

In this study, we employ the Invalsi-ITA benchmark to evaluate various Bambi models, a series of Italian BabyLMs which differ from one another in terms of i.) the amount of training data, ii.) the type of training data and learning strategies adopted, and iii.) instruction tuning (cf. Section 3.1). This benchmark is particularly well-suited to our analysis, as it allows us to observe improvements or declines across school grades and to isolate which of the above three variables may be influencing such trends in performance.

3. Evaluating Text Comprehension and Grammatical Knowledge with Invalsi-ITA

3.1. Models

The **Bambi** model is based on a lightweight GPT-2-style decoder architecture, with approximately 136 million parameters (Table 1). It is trained on a dataset composed of **transcripts of child-directed speech and multimedia content designed for children** [2]. So far, the dataset is organized into three tiers of increasing linguistic complexity, corresponding to the age ranges 0–6, 6–12, and 12–18. An additional tier is currently in progress. For the Bambi baseline model, all three tiers are used in a fully shuffled format. In contrast, the **Bambi_CL** (Curriculum Learning) model is trained on the tiers sequentially, progressing from the simplest to the most complex. Based on both the base and CL models, **Instruction Tuning** (IT) variants are implemented (Table 2). The IT training dataset comprises the following resources:

- `teelinsan/camoscio_cleaned` : a translated version [20] of the Stanford Alpaca dataset [21], which consists of LM-generated instruction-response pairs based on a seed set of human-written prompts [22]. The dataset contains approximately 50,000 items.
- `massimilianowosz/gsm8k-it` : a translated version of GSM8K [23], a dataset of 8.500 grade school-level math word problems.
- `Mattimax/DATA-AI_Conversation_ITA` : a dataset of Italian-language conversations, comprising 10,000 items [24].

For comparison purposes, the same architecture was trained on a traditional dataset of equivalent size, using a random subset of **mC4** [25], a corpus derived from the public Common Crawl web scrape and used to train standard LMs.

It is important to note that BabyLMs typically operate with limited input and output context windows, both to maintain model compactness and to respect cognitive plausibility constraints. In particular, the training data for the first and second developmental tiers avoid excessively long sequences. However, to enable evaluation on the Invalsi-ITA benchmark, the models were trained with a context window of 6,144 tokens, the minimum required to avoid truncating benchmark items. Crucially, our dataset remains untouched. The BabyLMs are compared against five other models (Tables 1 and 2). **Minerva-3B** is the model trained on the least amount of data, despite not being the smallest in size. It is followed by **Minerva-7B** and **Minerva 7B-it**, which rank second in terms of data volume [26]. Next is **Velvet-2B**, trained on approximately 3

Architecture	Vocabulary Size	Layers x Heads	Hidden Size	Trainable Parameters
Bambi	30,000	12x12	768	135,856,128
Minerva-3B	32,768	32x32	2,560	2,894,236,160
Minerva-7B	51,200	32x32	4,096	7,399,018,496
Velvet-2B	126,976	28x32	2,048	2,223,097,856
Cerbero-7B	32,000	32x32	4,096	7,241,732,096

Table 1
Hyperparameters of the models used in the experiment.

Model	Data size	Epochs	Curriculum Learning	Instruction Tuning
Bambi	86M words	16	no	no
Bambi_it	86M words	16	no	yes
Bambi_CL	86M words	[13,18,10]	3 steps	no
Bambi_CL_it	86M words	[13,18,10]	3 steps	yes
Bambi_mc4	86M words	20	no	no
Bambi_mc4_it	86M words	20	no	yes
Minerva-3B	660B tokens	1	no	no
Minerva-7B	2.48T tokens	1	no	no
Minerva-7B-it	2.48T tokens	1	no	yes
Velvet-2B	3T tokens	1	no	yes
Cerbero-7B	UNK	1	no	yes

Table 2
Training details of the BAMBI familiy models and the baseline models.

trillion tokens [1], and finally **Cerbero-7B**, for which the amount of training data has not been disclosed by the developers [27]. These models were chosen because their training corpora are predominantly in Italian.

3.2. Invalsi-ITA

Invalsi-ITA [3] is a benchmark derived from the adaptation of an established battery of assessments aimed at gauging educational proficiency throughout Italy.

The INVALSI (*Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione* 'National Institute for the Evaluation of the Education and Training System') tests have been administered to Italian students since the 2005/2006 school year. These tests are designed to monitor the students' competence of Italian language and Mathematics throughout their educational path. Increasingly complex tests are administered during primary school (grades 2 and 5), middle school (grades 6 and 8) and high school (grades 10 and 13).

[1] https://huggingface.co/Almawave/Velvet-2B

Invalsi-ITA focuses on the Italian language. It originally included 1,264 questions, classified by [3] into: i.) multiple choice; ii.) binary (e.g., TRUE/FALSE); iii.) open-ended; iv.) other. The authors of the benchmark excluded categories (iii.) and (iv.) retaining only multiple choice (87.47%) and binary (14.33%) questions, for a total of 1,117 questions. The benchmark assesses two main kinds of competence: **text comprehension** and **linguistic knowledge**. Text comprehension items (930/1,117, 83.26% of the total) require students to read a text and answer related questions (e.g., *Le prime tre righe del racconto parlano della vita di Polipetto nel suo ambiente. Quale frase spiega in poche parole come viveva Polipetto?* 'The first three lines of the story talk about Polipetto's life in his environment. Which sentence briefly explains how Polipetto lived?'), while language items (187/1,117, 16.74% of the total) assess knowledge of specific grammatical rules (e.g., *Indica in quale frase la parola "pietra" è usata in senso figurato, cioè non indica la pietra vera e propria.* 'Indicate in which sentence the word "stone" is used figuratively, that is, it does not refer to an actual stone.').

Question Macro-Area	Grade 2	Grade 5	Grade 6	Grade 8	Grade 10	Grade 13
Comprehension	149	275	58	245	190	13
Semantics	1	8	0	14	8	7
Syntax	0	27	7	35	18	7
Morphology	0	9	2	6	11	0
Phonology	0	3	0	1	0	0
Pragmatics/Textuality	0	0	0	1	0	0
Punctuation/Spelling	1	5	0	9	1	6
Total	151	327	67	311	228	33

Table 3
Internal structure of the Invalsi-ITA benchmark.

Table 3 summarizes the macro-areas covered by the questions in each grade (for more details, see [3, 19]).[2] Evaluating language models brings to the fore important questions about data contamination. The Bambi model was trained on a dataset specifically built and curated by the authors, ensuring it is free from contamination. For other models, verification is more challenging. Minerva models stand out for their transparency in this regard, and it appears safe to assume they were not exposed to the benchmark data. Cerbero-7B was released prior to the benchmark (2023 vs. 2024), so contamination also seems unlikely. Velvet-2B is more recent and its training dataset has not been made publicly available, making it difficult to assess potential overlap.

3.3. Method

The items are presented to the models in a zero-shot setting. Each item consists of a *text* (when present), a *question* that includes the list of multiple-choice options, and the *answer*, often represented only by the letter corresponding to the correct choice. Prompts and expected outputs are formatted using the following template (originally in Italian; a translation is provided here for clarity).

Prompt:

> Read the text and answer the question:
> `{text}`
> `{question}`

Completions:

- La risposta corretta è A: `{answer_a}`
- La risposta corretta è B: `{answer_b}`
- La risposta corretta è C: `{answer_c}`
- La risposta corretta è D: `{answer_d}`

A likelihood-based method was used to select the model's responses. Each model was presented with the prompt and the set of possible completions. The selected answer corresponds to the prompt–completion pair with the highest likelihood.

3.4. Results

Figure 1 shows the accuracy obtained by all models in each grade, considering both the text comprehension and the linguistic items. The accuracy values for each model in each grade are reported in Table 4 (Appendix 4).

A similar accuracy pattern emerges across grades 2 to 10 (Figure 1,). Cerbero-7B consistently achieves the highest accuracy, although its performance gradually declines over the grades. Minerva-7B and Minerva-7B-it follow with slightly lower scores, showing peaks in grades 2 and 6, a pattern also observed in Velvet-2B. In contrast, Minerva-3B aligns more closely with the Bambi models, which display the lowest accuracy throughout these grades.

A different pattern emerges in grade 13: Bambi, Bambi_it, and Bambi_mc4_it achieve the highest accuracy, alongside Velvet-2B. Slightly lower scores are obtained by the Minerva models, with Minerva-7B-it still leading this group. Notably, Cerbero-7B's performance drops significantly in this final grade. Focusing on the Bambi family, the strongest performances are overall exhibited by Bambi, Bambi_it, Bambi_CL_it, and Bambi_mc4_it.

Let us now turn to the accuracy the models achieved in the text comprehension items, displayed in Figure 2. The accuracy values are reported in Table 5 (Appendix A). The figure shows that the accuracy values and patterns observed for the comprehension items largely reflect those found in the overall analysis. Cerbero-7B consistently

[2] Due to the limited number of items within each linguistic macro-area, we opted to group all linguistic items together for the analysis. As a result, only comprehension and language items are discussed in Section 3.4.

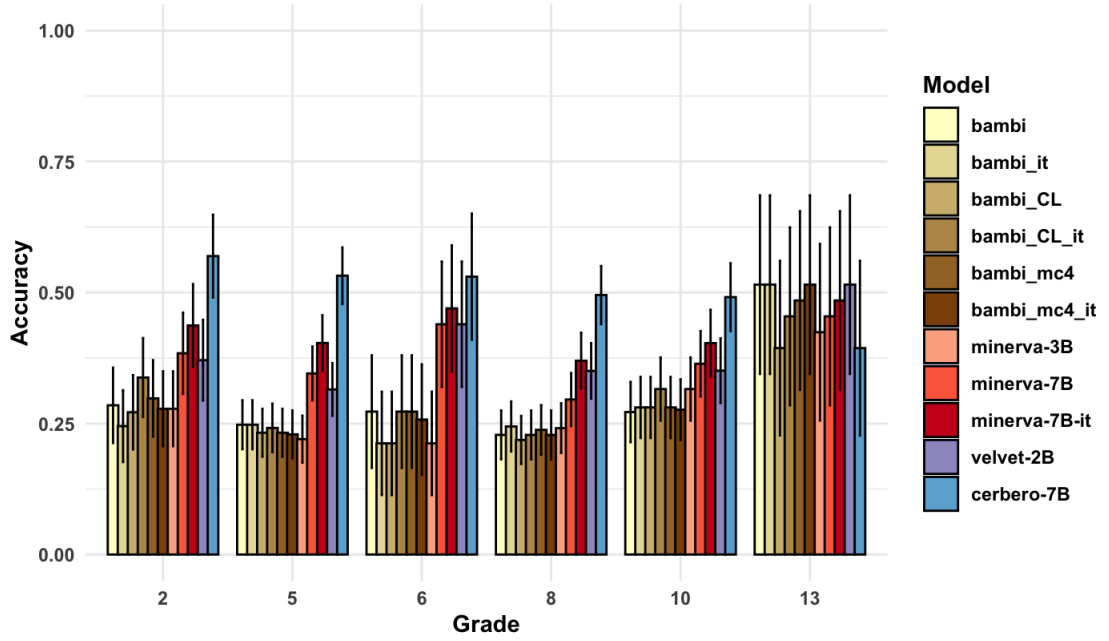

Figure 1: Accuracy reached by each model, for each grade, considering both the comprehension and the linguistic items. Error bars represent 95% confidence intervals.

achieves the highest accuracy across grades 2 to 10 (with all values above 0.50, though gradually declining), while a marked drop is observed in grade 13. Across grades 2 to 10, the Minerva models attain the second-highest accuracy, with Minerva-7B-it performing best within the family, closely followed by Minerva-7B. As in the overall analysis, the Bambi models perform poorly from grades 2 to 10 but improve significantly in grade 13: Bambi, Bambi_it, and Bambi_mc4_it all exceed 0.50 accuracy in this grade. The same pattern is observed for Velvet-2B.

A different trend is observed when considering only the accuracy achieved with respect to language items, displayed in Figure 3. The accuracy values are reported in Table 6 (Appendix A). Cerbero-7B, Velvet-2B, and Minerva-3B perform overall worse with respect to items specifically targeting grammatical knowledge than they do in text comprehension items. Minerva-7B and Minerva-7B-it, on the contrary, achieve similar accuracies in both tasks, and perform better in this task in grades 2 and 6. As for Bambi models, they differ from each other regarding the accuracy they achieve. In grade 2, only Bambi, Bambi_mc4, and Bambi_mc4_it achieve the highest accuracy (0.50) of all grades, whereas the others do not provide any correct answer in this grade. In grade 5 the same three Bambi models perform slightly better than Minerva-3B and Velvet-2B. In grade 6 Bambi_CL and Bambi_CL_it reach a peak in accuracy exceeding 0.50, followed by Bambi_mc4_it. Overall, grades 2 and 6 appear to be easier for some models, but challenging for others. Grade 13 is challenging for all models, as none of them provide a correct response.

Finally, let us take a look at the accuracy achieved by the models in the two kinds of questions that compose the Invalsi-ITA benchmark, i.e., multiple choice and binary (a summary of the accuracy values achieved for binary and multiple choice questions is reported in Table 7, given in Appendix A). The accuracies achieved for the binary questions are displayed in Figure 4.

For binary questions, accuracy generally hovers around or slightly above the expected chance level (0.5). Most models tend to perform better at the lower (grade 2) and upper (grade 13) ends of the evaluation spectrum, with a noticeable dip in performance across intermediate grades (5–10). Among the best-performing models, Bambi_CL_it and Cerbero-7B achieve the highest accuracy at grade 2 (0.70 and 0.65, respectively). Minerva-7B-it and Cerbero-7B show relatively stable performance across grade levels, with only minor fluctuations. Notably, Bambi_CL_it performs comparably to larger models.

Multiple choice questions (Figure 5) appear to be more challenging for all models. Given the four-alternative

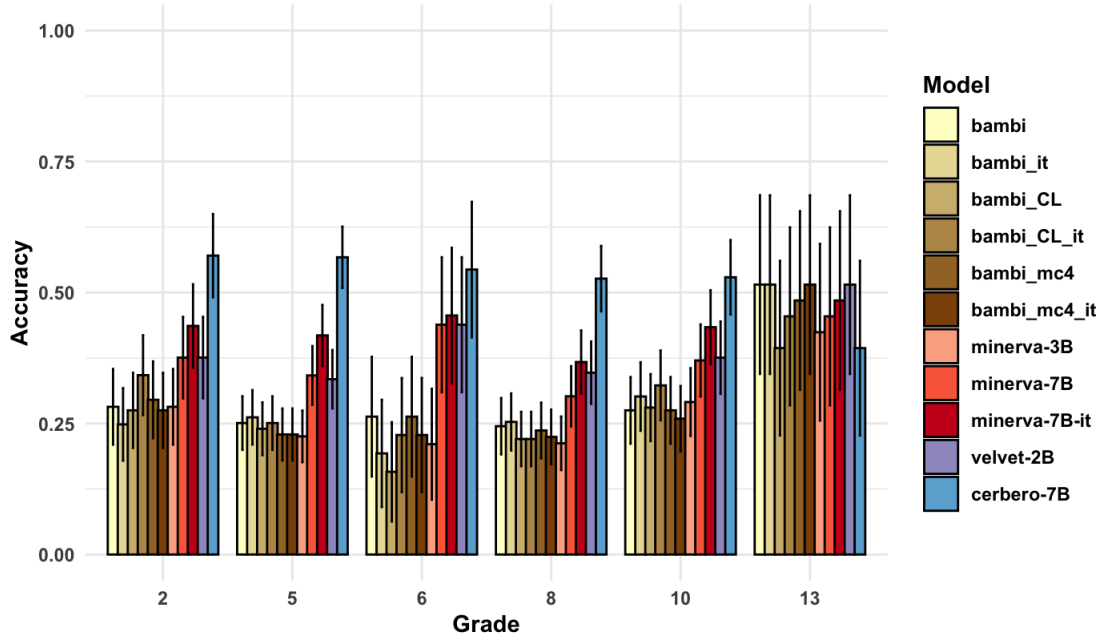

Figure 2: Accuracy reached by each model, for each grade, considering the text comprehension along with 95% confidence intervals (shown as error bars).

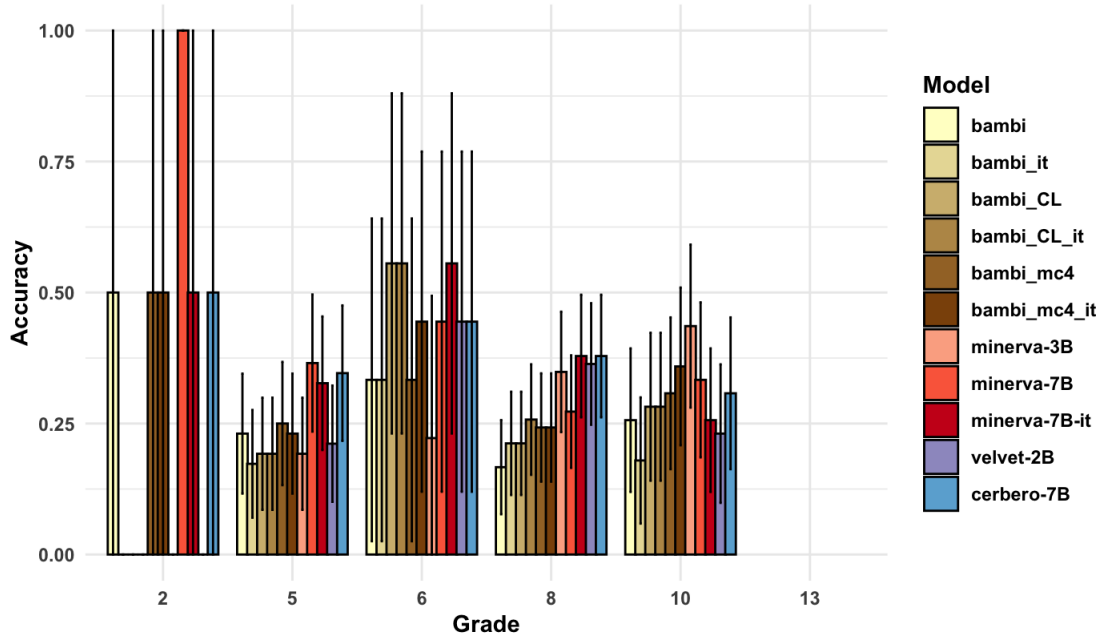

Figure 3: Accuracy reached by each model, for each grade, considering the language items along with 95% confidence intervals (shown as error bars).

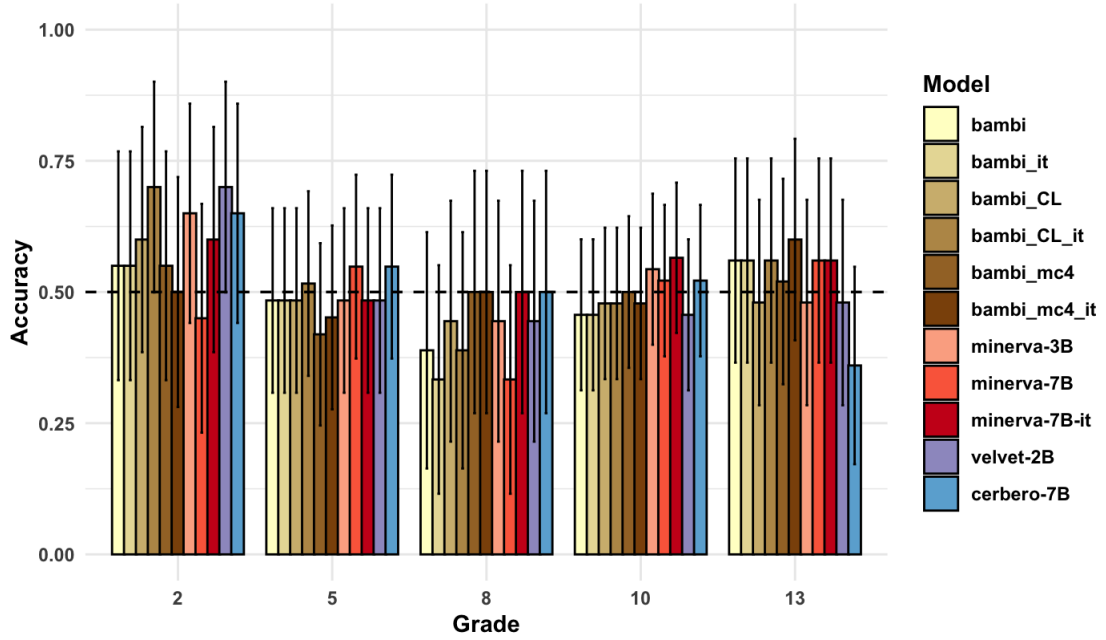

Figure 4: Accuracy achieved by all models in all grades with respect to binary questions along with 95% confidence intervals (shown as error bars). The dashed line represents the expected performance under random chance.

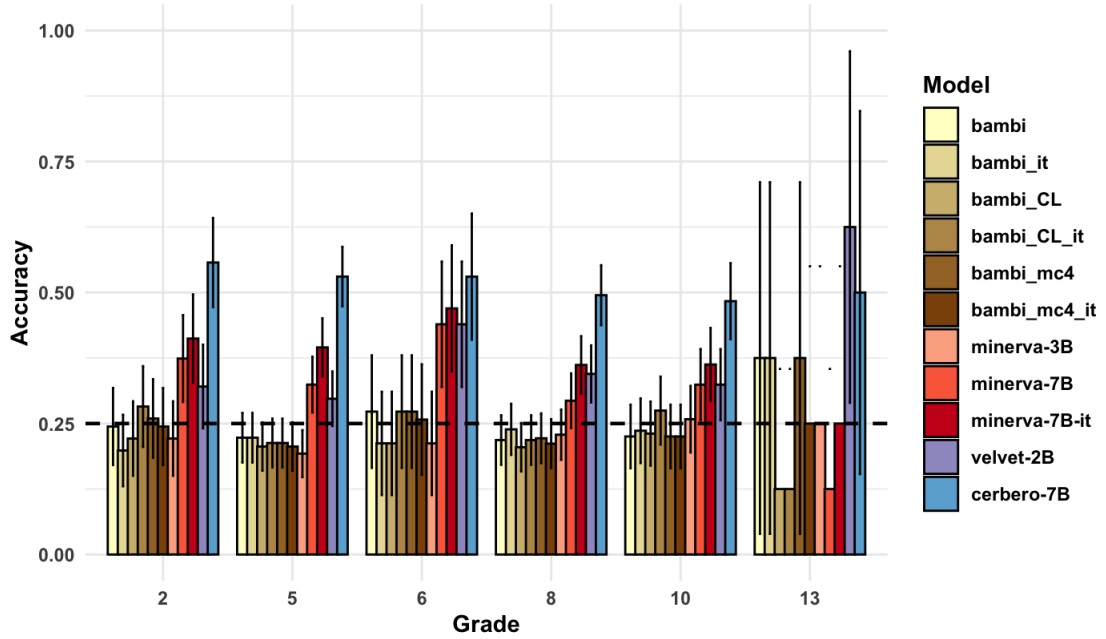

Figure 5: Accuracy achieved by all models in all grades with respect to multiple choice questions, along with 95% confidence intervals (shown as error bars). The dashed line represents the expected performance under random chance.

format, chance accuracy is approximately 0.25, and most models perform only marginally above this baseline. Still, some models demonstrate steady improvement across grade levels, particularly Velvet-2B and Cerbero-7B. The latter stands out as the most consistent and accurate performer in this task, achieving scores in the range 0.53 to 0.56 across several grades and peaking at 0.625 in grade 13. Bambi models, on the contrary, seem to find this kind of questions more challenging, particularly considering grades 2 to 10. However, Bambi, Bambi_CL_it, and Bambi_mc4 exceed the above-chance level in various grades. In particular, the performance of Bambi, Bambi_it, and Bambi_mc4 peaks at grade 13, reaching an accuracy around 0.40.

3.5. Discussion

The Invalsi-ITA benchmark appears to be challenging for all the models under investigation, as none of them exceed an accuracy value of 0.60. It should be kept in mind, however, that Invalsi tests are also challenging for Italian students [3]. [3].

The larger models, i.e., Cerbero-7B, Minerva-7B and Minerva-7B-it, perform overall better in this benchmark, especially when they are instruction-tuned. The reason may lie in the nature of Invalsi-ITA. This benchmark consists indeed of text comprehension items and language items, which specifically address normative grammatical rules, instead of the models' linguistic competence *tout-court*. Naturally, models which are exposed to a larger amount of training data and, even more importantly, to a large amount of *written* data, may be facilitated in these kinds of tasks, either because they have been exposed to the actual texts used in the benchmark, or because they are more used to this kind of linguistic input.

Nonetheless, Bambi models exhibit a great improvement in grade 13 with respect to the text comprehension items, and some of them perform comparably to larger models with respect to language items (e.g, in grades 2 and 6). These results suggest that compact models, despite lacking comprehensive world knowledge, can develop robust grammatical knowledge at early stages of training. Furthermore, considering binary questions, most of them, particularly Bambi_CL_it, Bambi_mc4 and Bambi_mc4_it, perform comparably to larger models in specific grades despite their compact size and training constraints, suggesting the potential benefits of a combination of oral and written training data.

Turning to curriculum learning and instruction tuning, a closer examination of the different Bambi models indicates that each strategy contributes modest gains, particularly in early grades. However, models that combine both strategies, such as Bambi_CL_it, show more consistent improvements, especially compared to IT-only variants. This is particularly evident in the case of the language items. The pattern implies that CL may enhance a model's capacity for subsequent learning, making IT more effective. This finding aligns with insights from human developmental learning, where structured progression lays the groundwork for improved adaptability and generalization over time [4].

These results give rise to some puzzling observations that merit closer examination. For instance, when comparing the Bambi models with their mc4-trained counterparts, substantial differences appear only in grades 2 (although this grade includes only two items) and 6 of the language items. This prompts the question of whether using ecologically plausible data is as crucial as often assumed, or if standard training corpora, such as mc4, can produce comparable results. In fact, the Bambi_mc4 models perform comparably to other Bambi models in many settings, indicating that the choice of data alone does not yeld substantial difference. However, they do not clearly outperform the Bambi models either: they achieve their best relative result in grade 5 of the language items, but in all other grades and tasks they perform worse or at best match the level of at least one of the Bambi variants. This pattern suggests that while web training data can approximate the results of carefully curated child-directed speech to some extent, it does not consistently provide an advantage, highlighting the need for a deeper analysis of the interactions between data quality, structure, and curriculum learning.

Another notable result is the unexpected jump in performance for the Bambi_CL models in grade 6 with respect to the language items. One possible explanation lies in the CL learning strategy: although the total number of tokens processed by these models over multiple epochs approaches the lifetime exposure of an 18-year-old adolescent, the absolute size of the Bambi dataset more closely reflects the typical linguistic input of a child aged six to eight. This alignment may account for the relatively strong results in grade 6, which corresponds to the final portion of the training curriculum. However, this interpretation does not readily explain another surprising outcome: in the text comprehension task for grade 13, the Bambi and Bambi_mc4 models outperform not only Bambi_CL and Bambi_CL_it, but also larger models like Minerva and Cerbero-7B. This could be an artifact of the limited number of items in this grade, but it highlights an area where further investigation is warranted to understand how data composition, curriculum

[3]Unfortunately, the benchmark does not provide student-level data. However, the paper describing the original resource [3] includes a bar plot illustrating the performance gap, which highlights the challenges faced by Italian students.

[4]We acknowledge the importance of cross-linguistic validation. To this end, we have submitted a related study to the third BabyLM Challenge [28], which is currently under review. Preliminary results on English show a similar trend.

pacing, and task type interact in shaping model behavior.

Taken together, these findings highlight several key insights. First, larger model size alone does not guarantee superior performance: smaller models can be competitive in specific cases, particularly in structurally simpler tasks. Second, apparently, training strategies such as CL and IT yeld effective improvements only under specific evaluation conditions. Finally, the performance gap between BabyLM and LLM remains substantial, particularly in tasks requiring semantic depth understanding or world knowledge. Closing this gap without compromising cognitive and linguistic plausibility remains a key challenge. Future work will need to explore new training strategies. and evaluation frameworks to address it.

4. Conclusion

In this work, we presented an evaluation of six Bambi model variants alongside five larger models, using the Invalsi-ITA benchmark, which assesses text comprehension and linguistic abilities.

This evaluation revealed that larger models are facilitated in the text comprehension task, because either they have already encountered the texts used in the benchmark or they are more used to this kind of linguistic input. Nonetheless, smaller but more cognitively plausible models appear to be facilitated in the learning and generalization processes, as highlighted by their improvement in higher grades considering both text comprehension and language items.

Acknowledgments

We acknowledge financial support under the PRIN 2022 Project Title "Computational and linguistic benchmarks for the study of verb argument structure" – CUP I53D23004050006 - Grant Assignment Decree No. 1016, 07/07/2023 by the Italian Ministry of University and Research (MUR), funded by the European Commission under the NextGeneration EU programme. This research was also partly funded by PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013—"FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme.

References

[1] L. Capone, A. Suozzi, G. E. Lebani, A. Lenci, et al., BaBIEs: A Benchmark for the Linguistic Evaluation of Italian Baby Language Models, in: Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), 2024.

[2] A. Suozzi, L. Capone, G. E. Lebani, A. Lenci, BAMBI: Developing BAby language Models for Italian, Lingue e linguaggio, Rivista semestrale (2025) 83–102.

[3] G. Puccetti, M. Cassese, A. Esuli, The invalsi benchmarks: measuring the linguistic and mathematical understanding of large language models in italian, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 6782–6797.

[4] E. G. Wilcox, M. Y. Hu, A. Mueller, A. Warstadt, L. Choshen, C. Zhuang, A. Williams, R. Cotterell, T. Linzen, Bigger is not always better: The importance of human-scale language modeling for psycholinguistics, Journal of Memory and Language 144 (2025) 104650.

[5] A. Warstadt, S. R. Bowman, What artificial neural networks can tell us about human language acquisition, in: Algebraic structures in natural language, 2022, pp. 17–60.

[6] A. Lenci, Understanding natural language understanding systems, Sistemi intelligenti 35 (2023) 277–302.

[7] L. Connell, D. Lynott, What can language models tell us about human cognition?, Current Directions in Psychological Science 33 (2024) 181–189.

[8] B.-D. Oh, W. Schuler, Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?, Transactions of the Association for Computational Linguistics 11 (2023) 336–350.

[9] A. De Varda, M. Marelli, Scaling in cognitive modelling: A multilingual approach to human reading times, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2023, pp. 139–149.

[10] A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, et al., Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora, in: Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning, 2023, pp. 1–34.

[11] M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, R. Cotterell, L. Choshen, A. Warstadt, E. G. Wilcox, Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora, in: The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, 2024, pp. 1–21.

[12] Y. Zhang, A. Warstadt, H.-S. Li, S. R. Bowman, When Do You Need Billions of Words of Pretraining Data?, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1112–1125.

[13] P. A. Huebner, E. Sulem, F. Cynthia, D. Roth, BabyBERTa: Learning more grammar with small-scale child-directed language, in: Proceedings of the 25th conference on computational natural language learning, 2021, pp. 624–646.

[14] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent Abilities of Large Language Models, Transactions on Machine Learning Research (2022).

[15] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, pp. 353–355.

[16] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, Advances in neural information processing systems 32 (2019).

[17] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, S. R. Bowman, BLiMP: The Benchmark of Linguistic Minimal Pairs for English, Transactions of the Association for Computational Linguistics 8 (2020) 377–392.

[18] L. Evanson, Y. Lakretz, J.-R. King, Language acquisition: do children and language models follow similar learning stages?, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 12205–12218.

[19] F. Mercorio, M. Mezzanzanica, D. Potertì, A. Serino, A. Seveso, Disce aut Deficere: Evaluating LLMs Proficiency on the INVALSI Italian Benchmark, arXiv preprint arXiv:2406.175352 (2024).

[20] A. Santilli, E. Rodolà, Camoscio: an Italian Instruction-tuned LLaMA, in: Proceedings of the Nineth Italian Conference on Computational Linguistics (CLiC-it 2023), 2023, pp. 385–395.

[21] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Alpaca: A strong, replicable instruction-following model, Stanford Center for Research on Foundation Models 3 (2023) 7.

[22] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 13484–13508.

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

[24] Mattimax, Italian conversations dataset by m.inc, 2025. URL: https://huggingface.co/datasets/Mattimax/DATA-AI_Conversation_ITA, dataset of over 10,000 prompt-response pairs in Italian, released by M.INC for training language models.

[25] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498.

[26] R. Orlando, L. Moroni, P.-L. H. Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, 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, pp. 707–719.

[27] F. A. Galatolo, M. G. Cimino, Cerbero-7B: A Leap Forward in Language-Specific LLMs Through Enhanced Chat Corpus Generation and Evaluation, arXiv preprint arXiv:2311.15698 (2023).

[28] L. Charpentier, L. Choshen, R. Cotterell, M. O. Gul, M. Hu, J. Jumelet, T. Linzen, J. Liu, A. Mueller, C. Ross, et al., BabyLM Turns 3: Call for papers for the 2025 BabyLM workshop, arXiv preprint arXiv:2502.10645 (2025).

A. Appendix A: Accuracy Values for Invalsi-ITA

Model	Grade 2	Grade 5	Grade 6	Grade 8	Grade 10	Grade 13
Bambi	0.28	0.25	0.27	0.23	0.27	0.51
Bambi_it	0.24	0.25	0.21	0.24	0.28	0.51
Bambi_CL	0.27	0.23	0.21	0.22	0.28	0.39
Bambi_CL_it	0.34	0.24	0.27	0.23	0.31	0.45
Bambi_mc4	0.30	0.23	0.25	0.24	0.28	0.48
Bambi_mc4_it	0.28	0.22	0.25	0.23	0.28	0.51
Minerva-3B	0.28	0.22	0.21	0.24	0.29	0.42
Minerva-7B	0.38	0.34	0.44	0.30	0.36	0.45
Minerva-7B-it	0.44	0.40	0.47	0.37	0.40	0.48
Velvet-2B	0.37	0.31	0.44	0.35	0.35	0.51
Cerbero-7B	0.57	0.53	0.53	0.49	0.49	0.39

Table 4
Accuracy achieved by each model in each grade, Invalsi-ITA (text comprehension and language items).

Model	Grade 2	Grade 5	Grade 6	Grade 8	Grade 10	Grade 13
Bambi	0.24	0.25	0.26	0.24	0.27	0.51
Bambi_it	0.25	0.26	0.19	0.25	0.30	0.51
Bambi_CL	0.27	0.24	0.16	0.22	0.28	0.39
Bambi_CL_it	0.34	0.25	0.23	0.22	0.32	0.45
Bambi_mc4	0.29	0.23	0.26	0.24	0.27	0.48
Bambi_mc4_it	0.27	0.23	0.23	0.22	0.26	0.51
Minerva-3B	0.28	0.22	0.21	0.21	0.29	0.42
Minerva-7B	0.37	0.34	0.44	0.30	0.37	0.45
Minerva-7B-it	0.43	0.41	0.46	0.37	0.43	0.48
Velvet-2B	0.37	0.33	0.44	0.35	0.37	0.51
Cerbero-7B	0.57	0.57	0.54	0.53	0.53	0.39

Table 5
Accuracy achieved by each model in each grade with respect to the text comprehension items.

Model	Grade 2	Grade 5	Grade 6	Grade 8	Grade 10	Grade 13
Bambi	0.50	0.23	0.33	0.16	0.26	0.00
Bambi_it	0.00	0.17	0.33	0.21	0.18	0.00
Bambi_CL	0.00	0.19	0.55	0.21	0.28	0.00
Bambi_CL_it	0.00	0.19	0.55	0.26	0.28	0.00
Bambi_mc4	0.50	0.25	0.33	0.24	0.31	0.00
Bambi_mc4_it	0.50	0.23	0.44	0.24	0.36	0.00
Minerva-3B	0.00	0.19	0.22	0.35	0.43	0.00
Minerva-7B	1.00	0.36	0.44	0.27	0.33	0.00
Minerva-7B-it	0.50	0.33	0.55	0.38	0.26	0.00
Velvet-2B	0.00	0.21	0.44	0.36	0.23	0.00
Cerbero-7B	0.50	0.35	0.44	0.38	0.31	0.00

Table 6
Accuracy achieved by each model in each grade with respect to the language items.

Model	Binary questions	Multiple choice questions
Bambi	0.49	0.25
Bambi_it	0.48	0.25
Bambi_CL	0.5	0.2
Bambi_CL_it	0.53	0.23
Bambi_mc4	0.5	0.26
Bambi_mc4_it	0.51	0.23
Minerva-3B	0.52	0.23
Minerva-7B	0.48	0.31
Minerva-7B-it	0.54	0.38
Velvet-2B	0.51	0.39
Cerbero-7B	0.52	0.52

Table 7
Summary of the accuracy reached by all models for binary and multiple choice questions.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) in order to: Grammar and spelling check and Formatting assistance. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.