BaBIEs: A Benchmark for the Linguistic Evaluation of Italian Baby Language Models

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

The possibility of comparing the linguistic competence of Language Models (LMs) to that of children has gained growing attention lately, raising the need for effective tools for evaluating both the former and the latter. To this purpose, we developed a resource for the linguistic evaluation of BabyLMs, which are LMs trained on datasets that comparable to the linguistic stimulus received by children. This resource adapts four standardized tests for the evaluation of linguistic skills of Italian-speaking children (BVL, TROG-2, TCGB-2 and Peabody). To verify the effectiveness of our benchmark, we administered it to Minerva, a LLM pretrained from scratch on Italian. Our results indicate that Minerva struggles to master certain linguistic aspects, achieving an age-equivalent score of 4 years, and that the type of task administered affects the model's performance.

Keywords

Language Models, Linguistic Evaluation, Benchmark, BabyLMs, Language Acquisition

1. Introduction

This paper presents BaBIEs (Baby Benchmark for Italian linguistic Evaluations), a new resource for the standardized evaluation of Italian BabyLMs, that is, language models (LMs) trained on datasets that are qualitatively and quantitatively comparable to the type of stimulus received by humans during language acquisition. The aim of this resource is twofold: (i) to evaluate the quality of the training data and strategies, in particular curriculum learning techniques, used in the development of BabyLMs and (ii) to provide a benchmark for comparing the performance of LMs, especially BabyLMs, with that of young human speakers. The paper is structured as follows: Section 2 reviews related work and delineates the rationale for this study; Section 3 details the characteristics of the BaBIEs benchmark, which results from the adaptation of standardized tests for evaluating the linguistic abilities of Italian-speaking children. In Section 4, we report a first test of the dataset with the Minerva Italian LM. The benchmark effectiveness is discussed in the light of the experiments in Section 5. Finally, in Section 6, some conclusions and possible future research directions are outlined.

2. Related works

2.1. Less is More

In recent years, LMs have progressively increased in both parameters number and volume of training dataset [1]. This trend presents several challenges, primarily (i) the escalating demand for data in the medium term could be a significant constraint on model development and enhancement [2]; and (ii) the mismatch between the volume and quality of training data for models and human learning behavior makes it difficult to compare their performance. This discrepancy poses methodological challenges for drawing conclusions or generalizations from studies of LMs in the context of language acquisition and cognitive modelling [3].

These challenges have spurred reflections on the relationship between the quantity and quality of training in natural language processing (NLP). Zhang et al. [4] address this topic by attempting to quantify the amount of text necessary for a LM to develop syntactic and semantic competence sufficient to achieve acceptable results in common NLP and natural language understanding (NLU) benchmarks. Specifically, the authors investigate the skills that can be acquired with training datasets ranging from 10 million to 100 million words. This range is derived from the well-known study by Hart and Risley [5]. According to them, a child is exposed to approximately 10 million words per year on average, reaching around

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

[*] Corresponding author.

[†] For the specific purposes of Italian Academy, Luca Capone is responsible for Sections 2 and 4.1, Alice Suozzi and Luca Capone for Section 3, Alice Suozzi for Sections 4.2 and 5, Alessandro Lenci and Gianluca E. Lebani for Sections 1 and 6.

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)

100 million words by age 10. Zhang et al. [4] demonstrate that substantial amounts of data are required to achieve good results in NLU tasks, such as those evaluated by SuperGLUE [6]. Performance improvements become noticeable after surpassing the threshold of 1 billion words and continue to improve steadily even beyond 30 billion words. However, tasks focusing on language syntax (e.g., acceptability judgment and minimal pairs) exhibit the most significant improvements between 1 million and 100 million words, after which the learning curve plateaus. The authors conclude that while acquiring factual knowledge necessitates large volumes of text, syntactic and semantic competence reaches saturation within the range of 10 million to 100 million words. Similar conclusions are reported by Wei et al. [7], who investigate the emergent skills of various LLMs, confirming that the most sophisticated behaviors primarily arise from scaling up model training. These findings justify the focus on BabyLMs, which are LMs trained on limited amounts of data, qualitatively resembling the stimuli received by a preschooler. Huebner et al. [8] illustrate this approach by training BabyBERTa on 50 million words of child-directed speech and simplified written text, achieving results comparable to RoBERTa-base on a grammar test suite. The BabyLM challenges [9] fall within this line of research, aiming to optimize model training through curriculum learning (CL) techniques and architectural optimizations. This approach not only makes research more affordable, but also results in models that are more cognitively plausible in comparison to human language acquisition. Although the proposed CL techniques did not lead to consistent improvements across all evaluation tasks [9], it has been demonstrated that a model trained with limited data (10 million words) can achieve results comparable to those of large LMs on various benchmarks.

2.2. Baby benchmarks for Baby models

These results prompt a reconsideration of the comparability between LMs training and human language learning. While benchmarks like BLiMP [10] and GLUE [11] facilitate comparisons between different models, they are not suitable for comparing BabyLMs to children who are acquiring a first language. Several studies attempt to address this shortcoming. For instance, Evanson et al. [12] compare the learning order of certain syntactic structures in English between GPT-2 and preschoolers. They find that the model exhibits a consistent order in learning syntactic structures, which aligns with the one observed in preschoolers. Other tests that compare training in LMs to human language acquisition include the reading time test [13] and the age-of-acquisition test [14].

For the Italian language, the three main benchmarks are: (i) UINAUIL [15], which includes six NLU tasks selected from the EVALITA (Evaluation campaign for Language Technology in Italian) archive; (ii) IT5 [16], which focuses on summarization tasks; (iii) the Invalsi benchmark [17], which evaluates the mathematical and linguistic competences of LMs in Italian. Only the latter is relevant to our study, as it allows a comparison between human language learning (in the school-age range 6-18 years) and that of the models. However, the age range considered by Invalsi involves more sophisticated NLU tasks, rather than the fundamental linguistic abilities learned during the preschool period, within the 100 million word budget.

3. Nurturing BaBIEs

In order to evaluate the linguistic abilities of BabyLMs, we developed BaBIEs by adapting four standardized tests designed to assess the linguistic competence of Italian-speaking children. These tests, which tap into different aspects of linguistic competence, are:

- *Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni (BVL)* 'Battery for the Assessment of Language in Children aged 4 to 12' [18]. BVL is designed to provide a global linguistic profile of Italian-speaking children and was standardized on a sample of 1,086 children aged 4 to 12. It consists of 18 tasks (e.g., semantic and phonological fluency, sentence and word comprehension, emotional prosody comprehension, etc.) grouped into three sections, i.e., production, comprehension, and repetition.

- *Peabody - Test di vocabolario recettivo* (Italian adaptation of the *Peabody Picture Vocabulary Test - Revised*) [19, 20]. PPVT-R is intended to measure the receptive vocabulary of the subject and was standardized on a sample of 2,400 aged 3 to 12 and 16. It consists of 175 items.

- *Test for Reception of Grammar - Version 2 (TROG-2)* [21]. TROG-2 is designed to assess the comprehension of verbal language, especially syntactic structures, and was standardized on a sample of 1,276 subjects aged 4 to 87. It consists of 20 blocks, each containing four items that focus on a grammatical structure (e.g., zero anaphor, reversible in and on, relative clause in object, etc.).

- *Test di Comprensione Grammaticale per Bambini - Seconda Edizione (TCGB-2)* 'Test of Grammatical Comprehension for Children - Second Edition' [22]. Analogously to TROG-2, TCGB-2 is a tool for assessing the comprehension of grammatical structures and was standardized on a sample of 455 children aged 4 to 11. It contains 74 items which measure the comprehension of six structures, i.e., the phenomenon of inflection, and five types of sentences: locative, active, passive, relative and dative.

It is worth noting that all tests are standardized on samples of typically-developing Italian-speaking subjects and are designed to be orally administered. That is, the stimuli are always read by the experimenter, and the child is asked either to answer orally or to point at a picture.

BaBIEs consists of five tasks (see Table 4 in Appendix A): this resource is twofold: (i) *Sentence Completion* (the only task assessing linguistic production), (ii) *Acceptability Judgment*, (iii) *Idiom Comprehension*, (iv) *Sentence Comprehension*, (v) *Lexical Comprehension.* These tasks are taken from BVL. We added 165 out of 175 items from Peabody (Lexical Comprehension task) and all the items contained in TROG-2 and TCGB-2 (both Sentence Comprehension tasks).[1] Except for the Sentence Completion task and the Acceptability Judgment task, all of the others are similarly-structured comprehension tasks. The child is presented with an oral linguistic stimulus (i.e., a word, a sentence or an idiom) and with a set of three or four possible answers, from which the child must choose the answer corresponding to the linguistic stimulus (the *target*). Together, a stimulus and its set of possible answers constitute a *test item.* The key factor in the process of item adaptation from the original tests to BaBIEs was the modality in which the sets of possible answers are displayed.

For the Acceptability Judgment task, we constructed minimal pairs of sentences by creating a grammatical or ungrammatical version of the verbal stimulus (depending on the (un)grammaticality of the original stimulus). In this task, the model receives one pair at a time. Its choice is determined by perplexity, with the sentence having the lowest perplexity score being chosen by the model.

For the Sentence Completion and Idiom Comprehension tasks, as both the stimuli and the sets of possible answers are linguistic expressions, the adaptation process only involved reformatting them to be readable by the model. The Sentence Completion task is modeled in a fill-in-the-blank format. The LM is given a textual sentence to complete, it receives one item at a time as input and generates up to three new tokens. The answer is considered correct if the correct completion appears in the generated sequence.

In contrast, the items for the Sentence and Lexical Comprehension tasks required substantial adaptation because these tasks involve pictures in their original version. The sets of possible answers are indeed presented on illustrated boards with four pictures, among which the child must choose the target picture that depicts the verbal stimulus. Adapting these items involved converting the pictures into linguistic expressions, either single words or complex sentences, which consist of the linguistic description of the distractor and target drawings. In the Sentence Comprehension task, the pictures were converted into sentences maintaining the lexical items constant whenever possible, and only altering the syntactic structure. This way, the target differs from the stimulus syntactically, but not lexically. For instance, given the linguistic stimulus *la pecora è spinta dal ragazzo* 'the sheep is pushed by the boy', the possible answers are: *cioè il ragazzo indica la pecora; cioè la pecora spinge il ragazzo; cioè il ragazzo spinge la pecora* (TARGET); *cioè il ragazzo guarda la pecora* 'that is, the boy indicates the sheep; that is, the sheep pushes the boy; that is, the boy pushes the sheep (TARGET); that is, the boy looks at the sheep'. Since the relevant structure is the reversible passive, target and distractors are active clauses with the same lexical items as the linguistic stimulus. For the Lexical Comprehension task, the converted target and distractors can be full sentences (especially if the stimulus is a verb), words, or phrases. Since the target converted from the target picture can not be identical to the stimulus word, we used a linguistic expression that is semantically-related to the stimulus (e.g., a synonym, hypernym, hyponym, etc.). For instance, given the stimulus *un trattore* 'a tractor', the set of possible answers is *cioè un microscopio; cioè una ruspa (TARGET); cioè un binocolo; cioè una bicicletta* 'that is, a microscope; that is, a bulldozer (TARGET); that is, binoculars; that is, a bicycle'. The target is *una ruspa* 'a bulldozer', which is semantically-related to the stimulus.

The adapted version of the Lexical Comprehension tasks (BVL and Peabody) functions as follows: each item comprises a textual lexical stimulus (a word) followed by a textual adaptation of the possible corresponding pictures, referred to hereafter as textual options (cf. Appendix A). The lexical stimulus is concatenated with each possible textual option to form four complex sentences. Noteworthy, we choose to concatenate the stimulus to each textual option by means of *cioè* 'that is', a conjunction used to clarify or restate something previously mentioned, which is particularly suited to make explicit the relationship between the the stimulus and the textual options. The model's choice is determined based on the perplexity obtained for each sentence. The same applies to the Sentence Comprehension tasks, which comprises items from the Sentence and Idiom Comprehension tasks (BVL, TROG-2, and TCGB-2). Some examples of adapted items (one per task) and the structure of the entire dataset are given in Appendix A.

[1] 10 out of 175 items from Peabody were excluded, because either the words were too rare to be known by BabyLMs, e.g., *emaciato* 'emaciated', or it was impossible to adapt the item without using visual stimuli, e.g., for *quadrato* 'square'.

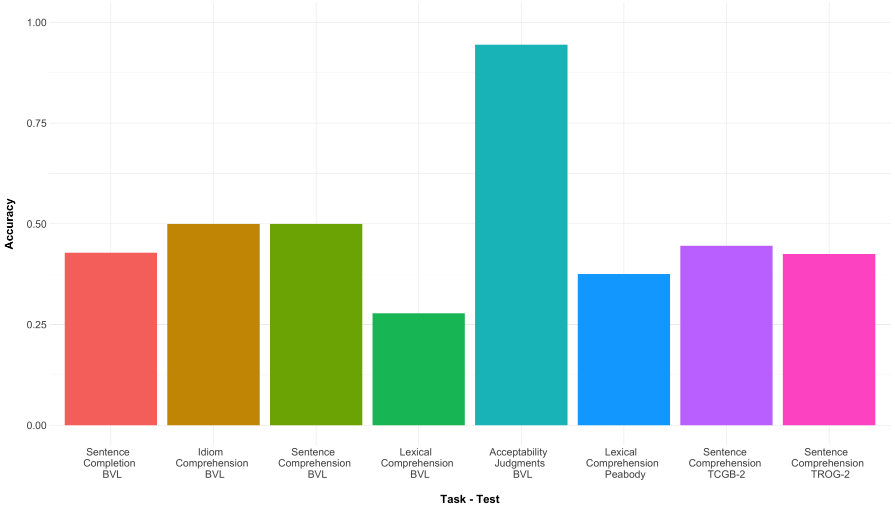

Figure 1: Accuracy obtained by Minerva in each task, across all tests.

4. Testing BaBIEs with Minerva

4.1. Model

To verify the effectiveness of this test, it was presented to a LM. Since no Italian LM primarily trained on child-directed speech and through curriculum learning was available, we opted for a conventional Italian LM[2]. Specifically, we chose `Minerva-3b-base-v1.0` (hereafter referred to as Minerva) [24], a decoder-only model (based on Mistral [25]) with 3 billion parameters. The choice was determined by the fact that, unlike other available models, Minerva was developed as an Italian model, despite also being pre-trained on a substantial amount of English text (660 billion tokens, 50% Italian and 50% English). For the experiments, the Huggingface implementation of the model was used. For the Sentence Completion task, we chose *beam search* as a generation strategy, with 3 beams. The models sampled the next generated token among the 50 most probable words. We combined this strategy with *nucleus sampling*, by setting a probability threshold of 0.95.

4.2. Results

The performance of Minerva is measured in terms of accuracy (number of true predictions relative to the total number of items). This measure is also used for evaluating children, allowing us to utilize standard scores to evaluate the model. The accuracy achieved by Minerva across all tasks is illustrated in Figure 1. Complete results, including accuracy for each clause type (Sentence Comprehension task - BVL, TROG-2, TCGB-2) and part-of-speech (Lexical Comprehension task - Peabody), are provided in Appendix B. Minerva obtains the highest accuracy in the Acceptability Judgment task (BVL) by far, with 17/18 true predictions and an accuracy of 0.94. Considering the standard scores, this falls between -1SD and +1SD for the age range 6.0-11,11 years (11,11 being the last age considered in the standardization of BVL). [3] The accuracy is lower for the Sentence Completion task (BVL), which - it is worth repeating - is the only production task, i.e., 0.43, with 6/14 true predictions. This score is positioned between -1SD and +1SD for the age range 4,0-5,5 years. In the Idiom Comprehension Task (BVL), the true predictions given by Minerva are 5/10, and the accuracy is of 0.5. This score is only seemingly low. Indeed, it falls between -1SD and +1SD for the age range 6,6-8,11 years and beyond +2SD for the age range 4,0-4,5 years. Let us now turn to the Sentence and Lexical Comprehension tasks (which involve picture-to-language conversion). We used three Sentence Comprehension tasks (from BVL, TCGB-2, TROG-2), which tap into partially different clause types (cf. Appendix B). In the BVL task, 20/40 true predictions are given by the model, corresponding to an accuracy of 0.5. The score is between -1SD and 0 for the age range 4,0-4,11 years. In the TCGB-2 task, the true predictions are 33/74, and the accuracy is 0.44.

[2] A new BabyLM [23] has been released a few weeks before the submission deadline. However, this model is not originally Italian but instead focuses on second language acquisition and its impact on the performance of a BabyLM.

[3] In standardized tests, the most frequent score obtained by children of a given age range is represented by 0. The typical range score extends from -2SD to +2SD from 0. For scores below -2SD, the performance is considered deficient. In this study, we consider the score range -1SD to +1SD, as we are not interested in potential language impairments.

According to the standard scores of TCGB-2, the model is placed between the 32nd and 45th percentiles for the age range 3,6-3,11 years. These percentiles correspond to the judgment of *within normal range* (as opposed to *excellent*, *good*, etc.) In the task adapted from TROG-2, Minerva reaches an accuracy of 0.42 (with 34/80 true predictions). In this test, the number of passed/failed blocks is relevant to the purposes of standard scores (a block being passed if the child provides the target response for at least 3/4 items). The model passes 6/20 blocks, obtaining an age-equivalent score of 4,1 years. The standard score for this age is 115, which falls into the 84th percentile. Finally, we used two Lexical Comprehension item sets (from BVL and Peabody). In the former (BVL), Minerva provides 5/18 true predictions, that correspond to an accuracy of 0.37. This score is below -2SD for the age range 4,0-4,5 years (4,0 years is the minimum age considered for the standardization). In the latter (Peabody), 62/165 predictions are true, the accuracy being 0.37. As mentioned above, we excluded 10 items from the adaptation process. Since the test age-equivalent scores are computed based on 175 items, we consider the raw-score range of 62-72 to establish the age-equivalent score of Minerva, so as to also take into account the excluded items. This raw-score range corresponds to the age-equivalent score range of 102-109 for the age range 3,9-4,2 years (i.e., between 0 and +1SD) and 92-99 for the age range 4,3-4,8 (i.e., between -1SD and 0).

5. Discussion

The scores obtained by Minerva generally align with the linguistic-age range 4.0-5.0. Variability in scores is observed i.) across different tasks, indicating that certain tasks may be easier for the model than others; and ii.) within the same type of task depending on the specific test they were adapted from (e.g., BVL–Sentence Comprehension, TROG-2). This discrepancy may be due to the adaptation of the test items, which, in turn, depends on the original distractor and target pictures. For instance, items in the Lexical Comprehension task of BVL required the model to make inferences to generate accurate predictions. Another possible factor (e.g., in the Sentence Comprehension task) is the complexity of specific syntactic structures evaluated by some tests. For instance, locative structures are particularly challenging for the model, as are passive clauses (cf. Appendix B). The model often fails to consistently grasp the rationale linking the stimulus and the target answer, likely due to Minerva not being an instruction-tuned model. Negation (Sentence Comprehension Task) is an illustrative example in this respect. BaBIEs contains 28 negative clauses (8/28 are passive clauses, and 20/28 are active clauses. Among the active clauses, 6 contain a double negation, i.e., *né...né* 'neither...nor'). Minerva selects the correct answer for 9/28 negative clauses (32.14%); of these, two are passives, six are active clauses, of which one contains a double negation. Wrong answers are selected for 19/29 negative clauses (67.86%), of which 6 are passives, 13 are active clauses, of which 5 containing a double negation. Four examples of wrong answers selected by Minerva are reported in Table 1. Such errors suggest that the model does not interpret negation, or in the case of clauses containing double negation, at least one of them, consistent with previous findings in the literature ([26], [27]). The complete sets of possible answers of the examples reported in Table 1) are given in Appendix C.

As can be seen in Table 1, the wrong answers selected by Minerva result from the failure to interpret the negation. In one case (i.e., the third example), the selected answer reveals that the model only interpreted the second (but not the first) negation.

The best score is obtained in the Acceptability Judgments task. This is not surprising and primarily due to the task being formulated with minimal pairs, a method proven to be particularly effective in testing LMs [10]. In the other tasks, the results are worse. Nonetheless, the age-equivalent score is not the whole story. In the Sentence Completion task, for instance, in spite of the low score obtained, the completions are not ungrammatical or nonsensical (cf. Table 2, more examples are provided in Appendix C). In the Lexical Comprehension tasks, the score further decreases. The results in both tasks (from BVL and Peabody) are fairly consistent, with an age score struggling to reach 4,5 years. The difficulties encountered by the model can be attributed to the limited context and the nature of the task, which is primarily semantic. The model also performs well in the Idiom Comprehension task, probably because idiomatic expressions are high-frequency expressions that a model trained on large amount of texts might easily have encountered. This could also explain why the score is lower for the Sentence Comprehension tasks, although the two are structurally similar. Indeed, unlike idiomatic expressions, the items of these tasks are less predictable and require a certain degree of inference for resolution, making their complexity more similar to that of Lexical Comprehension tasks.

6. Conclusions and future work

This paper presents BaBIEs, a novel resource specifically designed to evaluate the linguistic competence of BabyLMs and compare them to those of children. After having detailed the sources and the creation process of this resource, we provided the procedure for testing the Minerva model with the resource itself. Finally, we presented and discussed the results the model's performance.

Table 1
Examples of negative clauses, target answer, wrong answers provided by Minerva

Clause	Clause Type	Target Answer	Wrong Answer
La bambina non corre ‘The girl does not run’	ACTIVE	La bambina è ferma ‘The girl is still’	La bambina sta correndo ‘The girl is running’
Il cestino non è stato svuotato ‘The bin has not been emptied’	PASSIVE	Il cestino è pieno ‘The bin is full’	Il bambino ha svuotato il cestino ‘The boy emptied the bin’
La ragazza non sta né indicando né correndo ‘The girl is neither pointing nor running’	DOUBLE NEGATION	La ragazza è ferma ‘The girl is still’	La ragazza indica ma non corre ‘The girl is pointing but not running’
La scatola non è né grande né gialla ‘The box is neither big nor yellow’	DOUBLE NEGATION	La scatola è piccola e bianca ‘The box is small and white’	La scatola è grande e gialla ‘The box is big and yellow’

Table 2
Examples of model prediction for the Sentence Completion task

Verbal Stimulus	Model Completion	Correct Answer
La bambina si lava. Le bambine si ‘The girl washes herself. The girls’	**lavano**.‘wash themselves’ lavavano. ‘were washing themselves’ **lavano**. ‘wash themselves’	**lavano** ‘wash themselves’
Il cavallo corre nel campo. I cavalli ‘The horse runs in the field. The horses’	non possono correre ‘can’t run’ non hanno una ‘don’t have a.F.S’ non possono andare ‘can’t go’	**corrono** ‘run’

Based on the presented findings, the resource appears a valuable tool for evaluating not only BabyLMs but LMs in general. The poor performance exhibited by Minerva underscores the gap between child language acquisition and current language model training. This highligths the necessity for modifying model training to better encode human language and, more generally, human linguistic competence.

Future work will involve a more systematic linguistic analysis of the model’s performance, together with a comprehensive error analysis and a comparison to adult Italian-speakers. Furthermore, it will involve the development of a multimodal version of the test, which will more closely reflect the original tests and allow the evaluation of multimodal BabyLMs. Additionally, a BabyLM trained exclusively with Italian child-directed speech will be developed and evaluated with both the standard and multimodal versions of the test.

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 adopted on 07/07/2023 by the Italian Ministry of University and Research (MUR). 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] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).

[2] P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, A. Ho, Will we run out of data? an analysis of the limits of scaling datasets in machine learning, arXiv preprint arXiv:2211.04325 (2022).

[3] A. Warstadt, S. R. Bowman, What artificial neural networks can tell us about human language acqui-

sition, in: S. Lappin, J.-P. Bernardy (Eds.), Algebraic structures in natural language, CRC Press, Boca Raton, 2022, pp. 17–60.

[4] 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.

[5] B. Hart, T. R. Risley, Meaningful differences in the everyday experience of young American children, Brookes, Baltimore, 1995.

[6] 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).

[7] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] L. Evanson, Y. Lakretz, J.-R. King, Language acquisition: do children and language models follow similar learning stages?, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 12205–12218.

[13] E. G. Wilcox, T. Pimentel, C. Meister, R. Cotterell, R. P. Levy, Testing the predictions of surprisal theory in 11 languages, Transactions of the Association for Computational Linguistics 11 (2023) 1451–1470.

[14] T. A. Chang, B. K. Bergen, Word acquisition in neural language models, Transactions of the Association for Computational Linguistics 10 (2022) 1–16.

[15] V. Basile, L. Bioglio, A. Bosca, C. Bosco, V. Patti, 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), 2023, pp. 348–356.

[16] G. Sarti, M. Nissim, It5: Text-to-text pretraining for italian language understanding and generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 9422–9433.

[17] A. Esuli, G. Puccetti, The Invalsi Benchmark: measuring Language Models Mathematical and Language understanding in Italian, arXiv preprint arXiv:2403.18697 (2024).

[18] A. Marini, Batteria per la Valutazione del Linguaggio in bambini dai 4 ai 12 anni, Giunti Psychometrics, Firenze, 2015.

[19] L. M. Dunn, L. M. Dunn, Peabody Picture Vocabulary Test - Revised, American Guidance Service, Minneapolis, 1981.

[20] G. Stella, C. Pizzioli, P. E. Tressoldi, Peabody - Test di vocabolario recettivo, Omega, Torino, 2000.

[21] D. V. Bishop, Test for Reception of Grammar - Version 2, Giunti Psychometrics, Firenze, 2009.

[22] A. Chilosi, S. Piazzalunga, L. Pfanner, P. Cipriani, Test di Comprensione Grammaticale per Bambini-Seconda Edizione, Hogrefe, Firenze, 2023.

[23] Z. Shen, A. Joshi, R.-C. Chen, BAMBINO-LM:(Bilingual-) Human-Inspired Continual Pretraining of BabyLM, arXiv preprint arXiv:2406.11418 (2024).

[24] R. Orlando, P.-L. H. Cabot, L. Moroni, S. Conia, E. Barba, R. Navigli, Minerva-3b-base-v1.0, huggingface.co/sapienzanlp/Minerva-3B-base-v1.0 (2024).

[25] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[26] A. Hosseini, S. Reddy, D. Bahdanau, R. D. Hjelm, A. Sordoni, A. Courville, Understanding by understanding not: Modeling negation in language models, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter

of the Association for Computational Linguistics: Human Language Technologies, Online, 2021, pp. 1301–1312.

[27] T. H. Truong, T. Baldwin, K. Verspoor, T. Cohn, Language models are not naysayers: an analysis of language models on negation benchmarks, in: A. Palmer, J. Camacho-collados (Eds.), Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), Toronto, Canada, 2023, pp. 101–114.

A. Appendix A: Examples of adapted items

Table 3
Examples of the adapted items

Task	Verbal Stimuli	Set of possible answers & **Target answer**
Sentence Completion	Marco apre la porta. Anche noi <mask> ‘Marco opens the door. We, as well, <mask>’	**apriamo** **‘open’**
Acceptability Judgment	1. La bimba è buona ‘The child.F is good.F 2. La bimba è buono ‘The child.F is good.M’	**grammaticale** **‘grammatical’**
Idiom Comprehension	Quella donna cerca un ago in un pagliaio ‘That woman is searching a needle in a haystack’	1. cioè quella donna cerca tra la paglia ‘that is, that woman is searching through the hay’ 2. cioè quella donna si punge con l’ago ‘that is, that woman is pricking herself with the needle’ **3. cioè quella donna cerca qualcosa** **che è molto difficile da trovare** **’that is, that woman is looking for** **something that is hard to find’**
Sentence Comprehension	il cane non è seguito dal gatto ‘The dog is not followed by the cat’	1. cioè il gatto segue il cane ‘that is, the cat follows the dog’ 2. cioè il gatto segue il topo, ‘that is, the cat follows the mouse’ 3. cioè il cane segue il topo e il gatto segue il cane ‘that is, the dog follows the cat and the cat follows the mouse’ **4. cioè il cane segue il gatto** **’that is, the dog follows the cat’**
Lexical Comprehension	un trattore ‘a tractor’	1. cioè un microscopio ‘that is, a microscope’ **2. cioè una ruspa ‘that is, a bulldozer’** 3. cioè un binocolo ‘that is, binoculars’ 4. cioè una bicicletta ’that is, a bicycle’

Table 4
Structure of the dataset

Task	Subtypes (structure / PoS)	Number of items
Sentence Completion	none	14
Total:	—	14
Acceptability Judgment	none	18
Total:	—	18
Idiom Comprehension	none	10
Total:	—	10
Sentence Comprehension	Double negation	2
	Agreement	9
	Adversative Active	2
	Clitic	4
	Negative Active	10
	Relative Active	14
	Reversible Active	5
	Reflexive Active	2
	Reversible Affirmative Passive	8
	Negative Passive	8
	Reversible Negative Passive	1
	Affirmative Active	10
	Dative	6
	Inflection	16
	Locative	12
	Affirmative Passive	10
	Two Elements	4
	Negative	4
	Reversible ‘in’ and ‘on’	4
	Three Elements	4
	Reversible SVO	4
	Four Elements	4
	Relative Clause in the Subject	4
	Not only X but Y	4
	Reversible ‘above’ and ‘below’	4
	Comparative/Absolute	4
	Zero Anaphor	4
	Pronoun Gender/Number	4
	Pronoun Binding	4
	Neither nor	4
	X but not Y	4
	Post-Modified Subject	4
	Singular/Plural Inflection	4
	Relative Clause in the Object	4
	Centre-Embedded Sentence	4
Total:	—	194
Lexical Comprehension	Noun	121
	Verb	27
	Adjective	35
Total:	—	183
Total number of items:	—	419

B. Appendix B: Complete Results

Table 5
Accuracy obtained by Minerva, Sentence Comprehension Task (BVL), for each grammatical construction.

Construction	Number of true Predictions	Total Number of items	Accuracy
Double negation	2	2	1.00
Agreement	6	9	0.67
Adversative Active	0	2	0.00
Clitic	3	4	0.75
Negative Active	1	4	0.25
Relative Active	3	5	0.60
Reversible Active	2	5	0.40
Reflexive Active	0	2	0.00
Reversible Affirmative Passive	2	4	0.50
Negative Passive	1	2	0.50
Reversible Negative Passive	0	1	0.00
Total	20	40	0.50

Table 6
Accuracy obtained by Minerva, Sentence Comprehension Task (TCGB-2), for each grammatical construction.

Construction	Number of true Predictions	Total Number of items	Accuracy
Affirmative Active	4	10	0.40
Negative Active	3	6	0.50
Dative	5	6	0.83
Inflection	8	16	0.50
Locative	3	12	0.25
Affirmative Passive	5	10	0.50
Negative Passive	1	6	0.17
Relative	4	8	0.50
Total	33	74	0.45

Table 7
Accuracy obtained by Minerva, Sentence Comprehension Task (TROG-2), for each grammatical construction.

Construction	Number of true Predictions	Total Number of items	Accuracy	Failed/Passed Block
Two elements	3	4	0.75	PASSED
Negative	2	4	0.50	FAILED
Reversible ‘in’ and ‘on’	1	4	0.25	PASSED
Three elements	3	4	0.75	PASSED
Reversible SVO	1	4	0.25	FAILED
Four elements	2	4	0.50	FAILED
Relative clause in the subject	1	4	0.25	FAILED
Not only X but also Y	1	4	0.25	FAILED
Reversible ‘above’ and ‘below’	0	4	0.00	FAILED
Comparative/Absolute	4	4	1.00	PASSED
Reversible Passive	3	4	0.75	PASSED
Zero Anaphor	1	4	0.25	FAILED
Pronoun Gender/Number	2	4	0.50	FAILED
Pronoun Binding	1	4	0.25	FAILED
Neither nor	0	4	0.00	FAILED
X but not Y	1	4	0.25	FAILED
Post-Modified Subject	1	4	0.25	FAILED
Singular/Plural Inflection	0	4	0.00	FAILED
Relative Clause in the Object	4	4	1.00	PASSED
Centre-Embedded Sentence	3	4	0.75	PASSED
Total	34	80	0.42	6 PASSED / 14 FAILED

Table 8
Accuracy obtained by Minerva, Lexical Comprehension Task (Peabody), for each Part of Speech.

Part of Speech	Number of true Predictions	Total Number of items	Accuracy
Noun	43	103	0.42
Verb	9	27	0.33
Adjective	10	35	0.28
Total	62	165	0.37

C. Appendix C: Examples of Target and Wrong Answers Provided by Minerva

Table 9
Examples of model prediction for the Sentence Completion task

Verbal Stimulus	Model Completion	Correct Answer
La mamma cucina. Le mamme ‘The mother cooks. The mothers’	**cucinano** ‘cook’ **cucinano** per ‘cook for’ **cucinano**, ‘cook,’	**cucinano** ‘cook’
La bambina si lava. Le bambine si ‘The girlwashes herself. The girls’	**lavano** ‘washes themselves’ si lavavano ‘were washing themselves’ **lavano**, ‘wash themselves,’	**lavano** ‘wash themselves’
Il cavallo corre nel campo. I cavalli ‘The horse runs in the field. The horses’	non possono correre ‘can’t run’ non hanno una ‘don’t have a.F.S’ non possono andare ‘can’t go’	**corrono** ‘run’
Marco apre la porta. Anche noi ‘Marco opens the door. We do too’	entriamo. ‘enter.’ entriamo in ‘enter in’ entriamo e ‘enter and’	**apriamo** ‘open’
Il bambino gioca con la palla. Anche gli altri bambini ‘The boy plays with the ball. The.M other boys do too’	stanno giocando con ‘are playing with’ stanno giocando ‘are playing’ vogliono giocare con ‘want to play with [it]’	**giocano** (play)
Il bambino ha pianto tutta la notte. Anche ora lui ‘The child.M cried all night. Even now he	**sta piangendo** ‘is crying’ **piange**. ‘cries.’ **piange**, ‘cries,’	**piange** ‘cries’ **sta piangendo** ‘is crying’
Il papà parte spesso per lavoro. Anche ieri il papà ‘Dad often leaves for work. Yesterday too dad	**è partito** per ‘left for’ **è partito**. ‘left.’ **è partito**. ‘left.’	**è partito** ‘left’ **partiva** ‘was leaving’
Si sporca sempre giocando a calcio. Anche la volta scorsa ‘[He] always gets dirty playing soccer. Last time too’	, quando la ‘, when the.F’ , quando è ‘, when [he/she/it] is’ , quando I ‘, when I’	**si è sporcato** ‘[he] got dirty’ **si sporcò** ‘[he] got dirty’
Lui si perde spesso nelle grandi città. Anche qui ‘He always gets lost in big cities. Here too’	, come a ‘, like in’ , a Roma ‘, in Rome’ , in provincia ‘, in a small town/in the suburbs’	**si è perso**‘[he] got lost’ **si perderà** ‘[he] is getting lost’

Table 10
Examples of wrong and target answers selected by the model in the Sentence Comprehension Task, negative clauses

Verbal Stimulus	Set of possible answers & **Target answer**	Answer selected by the model
La bambina non corre 'The girl does not run'	1. La bambina sta correndo 'The girl is running' 2. Le bambine stanno correndo 'The girls are running' 3. La bambina raggiunge la mamma 'The girl reaches her mom' **4. La bambina è ferma** **'The girl is still'**	1. La bambina sta correndo 'The girl is running' (WRONG)
Il cestino non è stato svuotato 'The bin has not been emptied'	1. Il cestino è vuoto 'The bin is empty' **2. Il cestino è pieno** **'The bin is full'** 3. La mamma svuota il cestino 'The mom empties the bin' 4. Il bambino ha svuotato il cestino 'The boy has emptied the bin'	4. Il bambino ha svuotato il cestino 'The boy has emptied the bin' (WRONG)
La ragazza non sta né indicando né correndo 'The girl is neither pointing nor running'	1. La ragazza corre ma non indica 'The girl is running but not pointing' **2. La ragazza è ferma** **'The girl is still'** 3. La ragazza corre e indica 'The girl is running and pointing' 4. La ragazza indica ma non corre 'The girl is pointing but not running'	4. La ragazza indica ma non corre 'The girl is pointing but not running' (WRONG)
La scatola non è né grande né gialla 'The box is neither big nor yellow'	**1. La scatola è piccola e bianca** **'The box is small and white'** 2. La scatola è grande e gialla 'The box is big and yellow' 3. La scatola è piccola e gialla 'The box is small and yellow' 4. La scatola è grande e bianca 'The box is big and white'	2. La scatola è grande e gialla 'The box is big and yellow' (WRONG)