BabyBabelLM:
A Multilingual Benchmark of Developmentally Plausible Training Data

Anonymous ACL submission

Abstract

We present **BabyBabelLM**, a multilingual collection of datasets modeling the language a person observes from birth until they acquire a native language. We curate developmentally plausible pretraining data aiming to cover the equivalent of 100M English words of content in each of 45 languages. We compile evaluation suites and train baseline models in each language. BabyBabelLM aims to facilitate multilingual pretraining and cognitive modeling.[1]

1 Introduction

The prevailing trend in language modeling research is to prioritize scaling, both in terms of model size and training data volume (Kaplan et al., 2020; Choshen et al., 2024). While this approach has led to significant advances in model performance, it neglects fundamental research questions about the nature of language learning (Wilcox et al., 2024). It disincentivizes work on data-efficient modeling, which, from a practical perspective, offers benefits in terms of efficiency and accessibility. From a theoretical perspective, it ignores the growing mismatch between human language acquisition and language model (LM) learning. From infancy to maturity, English learners acquire language through exposure to less than 100M words (Gilkerson et al., 2017), several orders of magnitude less than the massive pretraining corpora required by contemporary LMs surpassing 10T words (Bengio et al., 2025).

In response to the field's focus on scale, the BabyLM Challenge (Warstadt et al., 2023) was created to redirect attention toward questions of data efficiency and developmental plausibility in language modeling. The shared task invites participants to propose data-efficient LMs pretrained on a fixed, developmentally plausible English corpus of child-directed speech (CDS), educational content, and other simplified texts. The top-performing submissions (Charpentier and Samuel, 2023, 2024) have significantly improved the state of the art for models trained on the same limited data budget, even surpassing LMs trained on much larger corpora on various benchmarks.

The BabyLM Challenge has generated a new line of research on data-efficient training and cognitively-inspired modeling (Warstadt et al., 2023; Hu et al., 2024), depending on the existence of developmentally plausible datasets as training corpora and supplying resources to ease and direct such research. However, the majority of this work has focused on English, largely due to the public availability of the pretraining corpora released for the BabyLM challenge, which is English-only. There is a small but growing body of work that extends the BabyLM research project beyond English (Salhan et al., 2024; Shen et al., 2024; Prévot et al., 2024; Matzopoulos et al., 2025; Padovani et al., 2025; Bunzeck et al., 2025). Such efforts are crucial for developing an accurate understanding of the relationship between human language acquisition and LM learning. Any claim that a model is developmentally plausible can only be truly substantiated by evaluations across typologically diverse languages, as there is variation in acquisition trajectories between languages, and human language learning also frequently occurs in multilingual settings (Grosjean, 1989; Slobin, 2014; Moran, 2016; Stoll, 2020).

To facilitate this research, we create BabyBabelLM, a multilingual collection of developmentally plausible training datasets. The collection includes 45 languages, encompassing families primarily rooted—though not exclusively spoken—in Europe, Asia, and Africa. For each language, we carefully select and compile publicly available datasets while prioritizing developmentally plausible data, as well as release new ones. This in-

[1] All code and data available through [redacted] and the supplementary materials.

cludes several categories of developmentally plausible data, such as CDS, educational resources, and other child-oriented content (e.g., books, news, and wikis aimed towards children). We sort languages into three tiers based on training set size, corresponding to the equivalent of respectively 100M, 10M, or 1M English words, calibrated by language-adjusted byte estimates (Arnett et al., 2024), to ensure comparability of data budgets across languages with differing orthographic and morphological characteristics.

To further facilitate research we also compile a list of evaluations to test models created on those domains. We provide a comprehensive list of existing datasets to facilitate any future questions and to provide coverage. Specifically, we cover both formal and functional competence across all languages, and include evaluation that fits the pretraining objective directly without adaptation (known as zero-shot), as well as fine-tuning based evaluation that relies on task-specific training datasets.

Overall, this effort releases:

- Developmentally plausible **pretraining datasets** for 45 languages, collected with licenses permitting research purposes (§3).
- A **pipeline** to allow for subsequent dataset expansion with new resources and languages (§3.3).
- A survey of multilingual **evaluation** tasks (§4) accompanied by an evaluation suite extendable by the community.
- A collection of 45 monolingual **pretrained models**, 7 bilingual models and a multilingual model that we analyze in §5.

2 Related Work

The first edition of the BabyLM challenge (Warstadt et al., 2023) released two pretraining corpora, respectively 10M and 100M words, each consisting of 39% developmentally plausible data and a selection of high-quality corpora (e.g. Wikipedia). The second edition (Hu et al., 2024) updated the datasets to increase the proportion of child-oriented data to 70%. Thus far, the BabyLM Challenge has been limited to English for training and evaluation. In both editions, BabyLM submissions were evaluated on two types of language tasks: 1) zero-shot minimal pair challenges (Warstadt et al., 2020; Ivanova et al., 2024) benchmarking linguistic competence, world knowledge or other capabilities by testing if the model prefers a correct sentence over an incorrect one with a minor but meaningful alteration; and 2) fine-tuning based evaluations where models are further trained on a novel dataset and tested on their ability to learn the underlying task.

Beyond English, a growing body of work has begun exploring BabyLM-style models and the collection of developmentally plausible training datasets for other languages. Salhan et al. (2024) propose acquisition-inspired curriculum learning strategies and train small-scale LMs on age-ordered CDS for French, German, Japanese, and Chinese. Prévot et al. (2024) investigate the value of spontaneous speech corpora for BabyLM evaluation with experiments on English and French. Matzopoulos et al. (2025) train BabyLMs for isiXhosa, a low-resource South African language, highlighting the limits of BabyLM research for languages without publicly available developmentally plausible corpora. Capone et al. (2024) release a corpus of Italian developmentally plausible training data. Padovani et al. (2025) show that training on CDS does not consistently improve grammatical learning across English, French, and German. Bunzeck et al. (2025) train LMs on distributionally varied subsets of a German BabyLM corpus, showing that syntax learning benefits from complex constructions while lexical learning benefits from fragmentary constructions. Finally, Shen et al. (2024) investigate developmentally plausible L2 acquisition by adapting an English BabyLM for Italian via a reward signal from a parent Italian model. However, these works are typically forced to compile novel datasets in addition to their scientific contribution, and they do not represent a coordinated effort to compile such training data in comparable ways and across diverse languages.

Some relevant multilingual resources do exist. Notably, the Child Language Data Exchange System (CHILDES; MacWhinney, 2000) is a multilingual database of transcribed child-adult interactions, including data for over 40 languages, with varying age ranges, interaction environments, and corpus sizes. CHILDES serves as a starting point for most of our languages. A previous effort to compile developmentally plausible multilingual training corpora is MAO-CHILDES (Yadavalli et al., 2023), an age-ordered dataset of CHILDES corpora for five typologically diverse languages (German, French, Polish, Indonesian, and Japanese), which is used to study cross-lingual training and L2 learning. Salhan et al. (2024) and Goriely and Buttery (2025) independently release MAO-CHILDES and IPA-CHILDES for four languages (Japanese, Chi-

nese, French, German) and a phonemized corpus based on CHILDES for 31 languages.

3 Dataset Creation and Overview

The BabyBabelLM dataset was created to support research on developmentally plausible language modeling across a wide range of languages. Our aim is to approximate the kind of linguistic input that humans are exposed to in early life, while providing clean, well-documented, high-quality data.

3.1 Data Collection Principles

The design of our datasets required various methodological choices regarding the types of data sources to include, ensuring their developmental plausibility and long-term extensibility. In this section, we describe the criteria guiding our choices, the organizational structure of our multilingual collection, and the licensing considerations.

3.1.1 Developmental Plausibility Criteria

Our guiding principle in dataset construction is that of **developmental plausibility**: the idea that pretraining data should approximate the linguistic input children encounter. To this end, we prioritized domains such as child-directed speech (CDS), educational materials, children's books, and transcribed conversations. We deliberately excluded synthetic corpora, like TinyStories (Eldan and Li, 2023) or TinyDialogues (Feng et al., 2024), despite their developmental intention, as synthetic data has been shown to feature a reduced long tail for many linguistic measures (Ju et al., 2025), more uniformity in syntactic constructions (Muñoz-Ortiz et al., 2024; Strübbe et al., 2025), and less alignment with human-like discursive patterns (Liu and Fourtassi, 2025). As such, it is unsuitable for our goal of approximating the full complexity of a child's linguistic environment.

In addition to content filtering, we prioritized data quality by removing noise, favoring conversational data when applicable, and standardizing the format of metadata (see Appendix A). This preserved realism enables controlled cross-lingual comparisons, which are essential for studying the impact of linguistic variation on model learning.

3.1.2 Community-driven Data Leadership

To ensure dataset quality, data collection for most languages was led by a researcher fluent in or familiar with that language. These language leads were responsible for sourcing appropriate corpora, verifying developmental plausibility, and coordinating with local experts in linguistics and acquisition.

The BabyBabelLM dataset is designed as a "living resource". As more developmentally plausible data becomes available, we aim to expand the collection both in breadth—by adding new languages—and in depth—by enriching existing ones. To support this, we provide an open-source pipeline that enables researchers to add entirely new languages and expand existing language datasets. While our initial release covers 45 languages, 16 languages rely solely on general-purpose multilingual data resources. We consider these entries as starting points for future, more comprehensive corpora.

We invite contributions through GitHub and Hugging Face, where researchers can submit new datasets, improvements, and evaluations. All additions are reviewed for compliance with our guidelines and incorporated into future versions of the dataset, ensuring proper attribution. We hope this model of open, collaborative development will lead to broader coverage and increased utility across diverse research agendas.

3.1.3 Licensing and Ethics

During our data collection effort, we verified that all data is released with licenses that permit academic research, such as Creative Commons or Public Domain. When licensing information was missing, the right holders of each source were contacted. We release our corpus with document-level licensing information and data source attribution to ensure that each resource is used ethically and within its rights. In the rare cases where no license or contact information was available, we decided to still release the data, but under a restrictive non-commercial license.

3.2 Dataset Composition

Constructing a multilingual dataset that is both developmentally plausible and broadly comparable requires careful attention to how data is organized. Languages differ widely in the availability and type of child-relevant resources, and these differences must be accounted for without undermining cross-linguistic consistency. This section outlines how we approached this challenge, the types of data we included, and how they differ from one another.

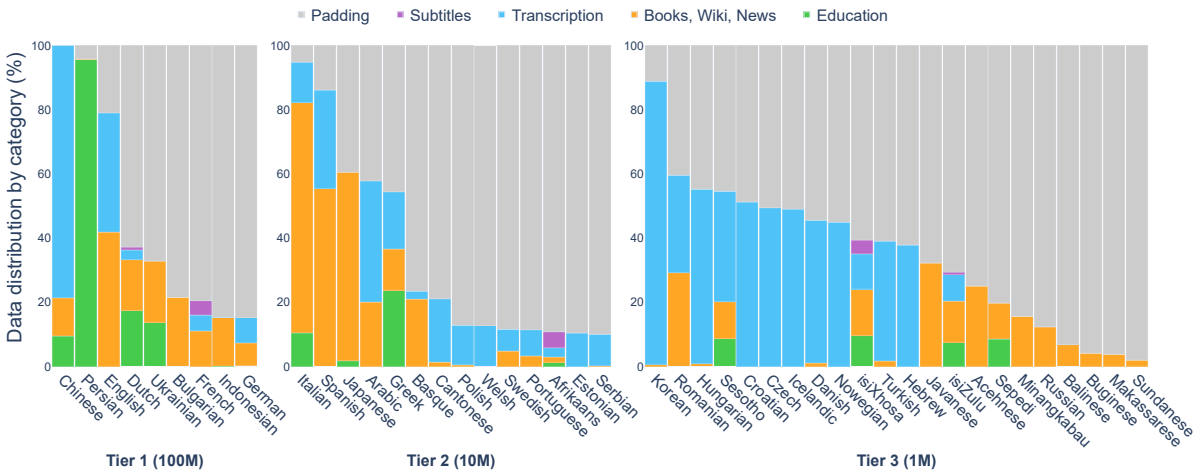

Figure 1: Training data distribution by category across languages for all data tiers in the BabyBabelLM dataset.

3.2.1 Data categories

Transcription Children learn language mainly from spoken input, which we therefore use as our primary data source. This child-directed speech (or *CDS*) differs drastically from the language data found in commonly used pretraining corpora. Usually, it is structurally short and simple (Genovese et al., 2020), and features high amounts of syntactic and lexical repetition (Tal et al., 2024), while its vocabulary is mostly restricted to everyday topics and children's immediate surroundings (Snow and Ferguson, 1977). The CHILDES database contains a large amount of such data in the form of recorded caretaker-child interactions (e.g., during free play, meal times, or shared book reading) and manually created transcriptions. We used all CDS available for our target languages in CHILDES as the base of our datasets. For some languages not written in Latin script (e.g., Japanese, Greek, Persian), the CHILDES data contains transliterations, this data is excluded from our data collection. As children also overhear language in their surroundings, we further included as much child-available speech (adult-adult dialogue) as possible per language.

Education We included educational content aimed at children, taken from textbooks and exams, as children spend a large amount of their time in education and encounter this kind of input regularly. On the content level, it provides much more direct instruction than CDS, which we deem useful for our purposes. After all, our BabyBabelLMs are not only supposed to learn formal linguistic patterns from the input, but ideally also more functional (visual semantic, pragmatic, and world) knowledge.

Books, Wiki, News To approximate the whole breadth of input that children receive, we further included child-oriented media, i.e., children's books, children's wikis, child-targeted news, and other appropriate media sources. For multilingual resources, we incorporated the Ririro story collection[2], GlotStoryBooks from Kargaran et al. (2023), and Child Wiki articles across many languages. Additionally, individual languages were enriched with monolingual resources. In contrast to CDS, this kind of data features longer and more complex sentences (Cameron-Faulkner and Noble, 2013; Bunzeck et al., 2025), and much more diverse vocabulary and content. As such, these sources should provide a useful training signal for more complex knowledge levels, similar to educational content.

Subtitles Finally, we also used movie/TV show subtitles suitable for children. While such fictional speech does differ from natural spoken data – for example, it features less hesitations, interjections, false starts or pauses (Bishop, 1991; Jucker, 2021; Gast et al., 2023) – it still approximates the linguistic properties of speech well and we deem it developmentally plausible. Furthermore, children are nowadays exposed to a wide variety of (video) media (Gowenlock et al., 2024), and thus encounter this kind of content regularly. We also include educational content subtitles from the QED corpus (Abdelali et al., 2014) for a small number of languages, filtering for data quality.

Padding To create comparable resources across languages, we pad our datasets to match the size

[2] https://ririro.com/

of different tiers (§3.2.2). For padding, we use the OpenSubtitles corpora (Lison and Tiedemann, 2016), which are significantly larger than our other data sources. To ensure our datasets do not contain content inappropriate for children, we omitted certain categories (e.g., adult content, crime, horror). For languages where not enough OpenSubtitles data is available, we further relied on FineWeb-C and Wikipedia data, among other resources, as fallback for additional padding.

3.2.2 Language Tiers and Coverage

Our dataset spans 45 languages drawn from a wide range of typological families. While Indo-European languages are well represented (22 out of 45), the collection also includes Semitic, Uralic, Bantu, Austronesian, and Sino-Tibetan languages, among others. This diversity was a key design goal, enabling investigation of language acquisition and modelling across distinct linguistic systems.

However, linguistic diversity is closely tied to disparities in data availability, resulting in big variations in data quantities for our set of languages. To enable fair comparisons, we classify languages into three distinct **tiers** according to the amount of collected data. Tier 1 includes languages with roughly 100 million English-equivalent tokens, Tier 2 with 10 million, and Tier 3 with 1 million. Ranking them by decreasing dataset size, the tiers contain 9, 15, and 21 languages respectively. This distribution further underscores the current scarcity of developmentally plausible corpora and the need for community-driven collection efforts.

Token thresholds are *calibrated* using the **byte premium** approach (Arnett et al., 2024), which adjusts for variation in orthographic and morphological structure by measuring the UTF-8 encoded size needed to express a fixed amount of content. For each language, we curated as much developmentally plausible content as possible before padding to the tier threshold using fallback data sources such as OpenSubtitles. Figure 1 summarizes the distribution of content categories across languages and tiers; more detailed per-language statistics are presented in Table 2.

3.3 Data Preprocessing

The data preprocessing is separated into two stages. Initially, language-specific preprocessing was carried out by the language leads, as needed by the specific data and language (more in Appendix B). Afterwards, we apply (and release) a uniform pipeline for all data, including standard normalization (unicode, whitespace, punctuation) and category-specific preprocessing. For dialogue transcripts, we remove linguistic annotationsFor subtitle data, we remove speaker labels, music note symbols, stage directions, and timestamps. For book-like formats (educational materials, children's books, wikis) and the QED dataset, we remove XML tags and URLs.

For language and script validation, we use GlotLID v3 (Kargaran et al., 2023). We classify sentence-like chunks of text, created by splitting documents into paragraphs and applying sentence-based heuristics. The document's final language is assigned via a segment-based majority vote. To maintain data quality, we filter mismatched segments within documents and discard any document that fails the overall validation. Other document metadata fields, such as the text category and license, are validated for the correct type and values when applicable (see Table 3).

4 Evaluation Suite

We create a multilingual evaluation suite that targets both the *formal linguistic competence* (knowledge of linguistic rules and patterns), and the *functional linguistic competence* (understanding and using language in the world) (Mahowald et al., 2024). We reviewed a large number of existing multilingual and monolingual benchmarks (Huang et al., 2025) with the aim of ensuring all our languages have at least one evaluation dataset testing formal and one testing functional linguistic competence.

Formal competence To assess formal linguistic competence, we prioritized high-quality, language-specific minimal pair benchmarks that target a diverse set of linguistic phenomena. This approach was applied to languages such as Basque (Kryvosheieva and Levy, 2025), Chinese (Liu et al., 2024), Japanese (Someya and Oseki, 2023), German (Vamvas and Sennrich, 2021), and Turkish (Başar et al., 2025). Where this was not possible, we employed datasets covering fewer phenomena but spanning multiple languages. In particular, for English, French, German, Russian, and Hebrew, we used CLAMS (Mueller et al., 2020), a cross-lingual minimal pair benchmark built from linguist-curated templates, focusing on subject-verb number agreement. In our experiments we refer to the collection of these tasks as *MonoBLiMP*. Finally, we incorporated

MultiBLiMP (Jumelet et al., 2025), a large-scale dataset of minimal pairs automatically generated from the Universal Dependencies treebanks (Nivre et al., 2017). MultiBLiMP targets subject-verb agreement in number, person, and gender, and offers the widest language coverage among our benchmarks, as detailed in Table 1.

Functional competence We include two types of benchmarks to evaluate functional competence. The first category focuses on factual and domain-specific knowledge memorized by the model, such as Global-MMLU (Singh et al., 2025), INCLUDE (Romanou et al., 2024), and BMLAMA (Qi et al., 2023). The second category assesses general reasoning abilities, including natural language inference, commonsense reasoning, narrative understanding, and reading comprehension. Benchmarks in this category include XNLI (Conneau et al., 2018), MultiNLI (Williams et al., 2018), HellaSwag (Zellers et al., 2019), Belebele (Bandarkar et al., 2024), ARC (Clark et al., 2018), xstorycloze (Lin et al., 2022b), TruthfulQA (Lin et al., 2022a), XCOPA (Ponti et al., 2020), SIB-200 (Adelani et al., 2024), and XWinograde (ai2, 2019; Cheng and Amiri, 2024). Additionally, we included XCOMPS (He et al., 2025), a multilingual conceptual minimal pair dataset with 17 languages.

Evaluation We evaluate these tasks in two ways. Tasks that are expressed as minimal pair comparisons are evaluated using **zero-shot** prompting, based on the model's output probabilities. For conducting these evaluations, we relied on Eleuther AI's LM Evaluation Harness (Gao et al., 2024). Tasks in this category are: all linguistic minimal pair tasks, XCOMPS, HellaSwag, Winogrande and XStoryCloze. For tasks involving classification and question answering we report performance after **finetuning**. The tasks on which we applied finetuning are: ARC, TruthfulQA, BMMLAMA, Belebele, INCLUDE, SIB-200, Global-MMLU, MultiNLI, XNLI and XCOPA. We initially experimented with zero-shot prompting on these tasks as well, but the limited data size of our corpora does not allow for in-context learning mechanisms to be acquired.[3] For finetuning, we adapt the pipeline from the *BabyLM Challenge 2024* evaluation framework[4]. We limit the number of training items to a max of 8,000 items, and finetune for 10 epochs using an 80/20 train/test split.

5 Experiments

Building on the resources outlined above, we train monolingual, bilingual and multilingual models to evaluate our benchmark suite.

Setup For training our models, we adopt the model configurations of the GoldFish model suite (Chang et al., 2024). For the monolingual models, we use a lightweight GPT-2 architecture with 4 transformer layers, 8 attention heads, and a hidden size of 512. The model uses GELU activations and standard dropout regularization (0.1) across attention, embeddings, and residual connections. It includes a feedforward inner dimension of 2048 and supports sequences up to 512 tokens. For all languages, we use a BPE tokenization (trained on the training corpus), with a vocabulary size of 8,192 tokens (Huebner et al., 2021). This results in small LMs of only 17.1M parameters. Each model is trained for 10 epochs.

For the bilingual models, we train a model using data from each language in Tier 1 and the English BabyLM (200M tokens total), keeping model configuration the same, but reducing training epochs to 5. For the multilingual model, we increase the number of layers to 12, hidden size to 768, and vocabulary size to 32,768, accommodating the wide range of languages and scripts this model should handle. The model is trained for only 1 epoch (around 1B tokens in total), and has 111M parameters. We additionally compare performance against Qwen3-0.6B (Yang et al., 2025), a capable multilingual LM of modest scale. Finally, we also experimented with training GPT-BERT (Charpentier and Samuel, 2024) models on our data, the architecture that has won the BabyLM challenge of 2024. Since these models did not outperform our GPT-2 models, we report their performance in Appendix C instead.

Results The results for our monolingual models are presented in Table 1. Linguistic benchmarks such as MultiBLiMP yield promising results, with Tier 1 models typically scoring above 80%. Performance on MultiBLiMP is strongly driven by data size, with Tier 2 and 3 languages performing worse. Performance on other benchmarks re-

[3] Olsson et al. (2022) show that the *induction heads* required for in-context learning develop only after exposure to 2.5-5 billion tokens. Developing sample-efficient methods that enable such mechanisms to emerge under much smaller data budgets, as targeted by the BabyLM challenge, is an important direction for future work but beyond the scope of this paper.

[4] `github.com/babylm/evaluation-pipeline-2024`

	Language	sib200	mnli	xnli	include	bmlama	global-mmlu	arc	truthfulqa	belebele	xcopa	multiblimp	monoblimp	xcomps	hellaswag	winogrande	xstorycloze
	Random	14.3	33.3	33.3	25.0	10.0	25.0	25.0	25.0	25.0	50.0	50.0	50.0	50.0	25.0	50.0	50.0
TIER 1 (100M)	Bulgarian	63.7	51.0	47.6	26.5	13.7	–	28.2	29.7	23.3	–	90.8	–	–	26.8	52.0	49.5
	Chinese	82.6	52.0	49.6	30.7	17.4	28.1	26.6	28.8	26.1	49.2	–	70.2	55.1	26.8	49.2	48.7
	Dutch	72.6	50.1	–	37.1	24.6	29.2	26.1	31.4	23.9	–	90.5	–	52.4	26.5	50.0	49.1
	English	75.1	50.1	49.1	–	15.6	28.7	27.0	22.9	30.0	47.3	82.0	65.9	–	26.5	51.4	49.5
	French	72.6	54.5	45.4	30.3	20.3	28.2	27.2	24.6	25.6	–	94.1	69.7	50.6	26.4	51.3	47.6
	German	65.2	51.7	48.2	51.6	15.2	27.5	28.0	25.4	26.1	–	88.6	77.1	52.6	25.9	51.8	48.8
	Indonesian	73.6	52.4	–	27.8	13.4	28.3	26.1	25.4	28.9	49.2	–	–	–	27.3	53.1	50.5
	Persian	80.1	50.6	–	31.3	15.8	29.0	28.8	28.0	27.2	–	71.3	–	53.6	26.4	50.7	52.2
	Ukrainian	71.1	48.9	–	32.7	14.7	28.3	25.3	27.1	23.3	–	88.6	–	50.6	26.4	50.3	47.6
TIER 2 (10M)	Afrikaans	71.6	52.6	–	–	14.8	–	28.6	28.8	28.8	–	–	–	–	26.1	51.3	49.5
	Arabic	45.8	44.3	37.1	36.8	15.3	28.5	27.8	28.0	25.6	–	75.9	–	52.9	25.8	47.4	46.8
	Basque	72.6	–	45.4[1]	30.7	–	–	24.7[1]	31.5[1]	24.4	49.2[1]	94.5	65.3	–	–	–	50.6
	Estonian	55.2	48.2	–	32.0	14.2	–	25.1	24.6	26.7	49.2	81.5	–	–	25.5	50.7	45.8
	Greek	58.2	48.8	46.1	34.7	13.0	26.1	28.8	21.2	26.1	–	89.2	–	50.3	26.4	49.5	49.4
	Hebrew	64.7	48.3	–	27.7	13.2	27.0	27.4	23.7	21.1	–	70.2	59.5	51.6	26.1	50.2	49.6
	Italian	61.7	50.5	–	31.9	14.6	28.3	27.0	28.0	21.1	55.0	77.5	–	–	26.5	50.1	50.2
	Japanese	67.7	44.6	–	26.0	11.9	27.9	28.0	27.1	24.4	–	–	61.9	50.8	25.1	47.5	47.8
	Polish	53.7	46.7	–	27.4	12.5	28.8	28.4	28.0	26.7	–	75.9	–	–	25.5	49.0	49.5
	Portuguese	61.7	49.7	–	28.2	14.3	28.7	25.9	21.2	26.1	–	80.7	–	–	26.3	48.8	48.8
	Serbian	37.3	41.3	–	38.0	13.4	28.7	25.3	31.4	27.8	–	–	–	–	25.6	49.5	48.5
	Spanish	66.7	50.5	49.0	34.2	15.7	28.2	26.3	26.3	23.9	–	83.0	–	51.3	26.4	49.1	47.8
	Swedish	55.2	48.4	–	–	13.9	26.7	27.2	23.7	27.2	–	100.0	–	–	25.9	49.3	48.1
	Welsh	73.6	–	–	–	–	–	–	–	–	–	91.4	–	–	–	–	–
	Yue Chinese	80.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
TIER 3 (1M)	Achinese	35.3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Balinese	52.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Buginese	29.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Croatian	40.3	37.8	–	26.5	12.7	–	24.5	27.1	26.7	–	–	–	–	25.8	47.6	45.6
	Czech	35.8	40.6	–	–	12.4	29.0	26.3	28.8	29.4	–	59.0	–	–	25.8	50.4	48.6
	Danish	42.3	45.9	–	–	15.3	–	26.1	23.7	27.2	–	82.0	–	–	25.7	49.0	48.9
	Hungarian	26.4	43.0	–	26.5	14.0	–	27.8	27.1	28.3	–	68.9	–	49.9	25.6	48.0	48.5
	Icelandic	41.8	46.0	–	–	11.8	–	26.1	27.1	30.0	–	71.6	–	–	25.4	50.0	45.7
	Javanese	53.7	48.3	–	–	12.7	–	26.8	26.3	21.7	–	–	–	–	25.8	50.5	50.6
	Korean	41.8	43.6	–	25.5	13.4	27.7	24.5	33.1	27.2	–	–	–	51.3	25.0	48.9	47.5
	Makasar	89.1[2]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Minangkabau	21.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Norwegian	49.2	47.3	–	–	12.0	–	28.6	24.6	25.6	–	–	–	–	25.9	47.0	47.1
	Sepedi	53.7	–	–	–	–	–	–	–	25.6	–	–	–	–	–	–	–
	Romanian	49.8	46.0	–	–	12.1	28.9	28.0	28.0	30.6	–	74.1	–	–	25.5	48.6	46.2
	Russian	43.3	44.6	44.0	37.0	12.5	29.1	27.4	28.0	31.1	–	58.5	52.0	49.1	25.9	51.3	48.8
	Sesotho	56.2	–	32.9[3]	–	–	26.2[3]	–	–	27.8	–	–	–	–	–	–	–
	Sundanese	62.7	–	–	–	–	–	–	–	26.7	–	–	–	–	–	–	–
	Turkish	34.8	36.7	37.5	31.6	14.3	28.6	30.7	25.4	25.0	49.2	64.9	59.8	51.6	25.9	49.9	50.1
	isiXhosa	46.3	–	32.4[3]	–	–	27.0[3]	–	–	23.9	–	–	–	–	–	–	–
	isiZulu	50.2	–	36.2[3]	–	–	30.3[3]	–	–	26.7	–	–	–	–	–	–	–
		FINETUNED										ZERO-SHOT					

Table 1: Performance of the monolingual models trained on BabyBabelLM. All scores denote average accuracy scores, either with 0-shot prompting on the base model or on the finetuned model. Zero-shot performance for all tasks is provided in Table 5. [1]For Basque we took XNLI, ARC, ThruthfulQA and XCOPA datasets from `HiTZ/xnli-eu`, `HiTZ/ARC-eu`, `HiTZ/truthfulqa-multi-MT` and `HiTZ/XCOPA-eu`. [2]For Makasar we used a similar task to SIB200 from `nusaparagraph_topic`. [3]For three South African languages (Sesotho, isiXhosa, isiZulu) we employed XNLI and Global-MMLU data from `afrixnli` and `afrimmlu`.

mains close to random chance (e.g., XCOPA, ARC, XCOMPS, HellaSwag). As such, our comparatively tiny BabyLMs only provide a starting point for further experimentation.

We further compare the results of our monolingual models for MultiBLiMP and Belebele to the multilingual BabyLM model (MultiBabyBabelLM) and to Qwen3-0.6B. Results are summarized in Figure 2; full results for Qwen are included in Table 4. On MultiBLiMP, the monolingual models generally outperform the multilingual one, except in four Tier 3 languages where the latter shows modest improvements. Compared to Qwen, results are mixed: our multilingual model is outperformed by Qwen in most cases, but remains stronger in eight languages, with no clear trend by tier. On Belebele, both our models perform near chance, while Qwen achieves substantially higher scores in all languages. This pattern extends to most other benchmarks, where Qwen consistently exceeds baseline and outperforms our models on knowledge-intensive and reasoning tasks.

Figure 3 presents results for the bilingual models, focusing on the three best- and worst-performing tasks from the monolingual models, along with SIB-200. All results are reported for zero-shot performance. For several tasks—SIB-200, BMLAMA, XCOMPS, and INCLUDE—adding English as a

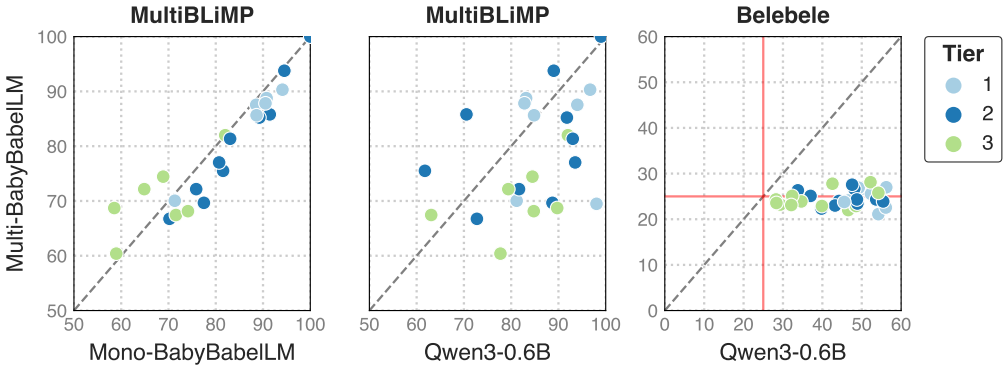

Figure 2: Language-level performance of the multilingual BabyBabelLM model against the monolingual models and Qwen3-0.6B on MultiBLiMP and Belebele. Each point denotes the accuracy on a specific language. Random performance for Belebele is denoted in red.

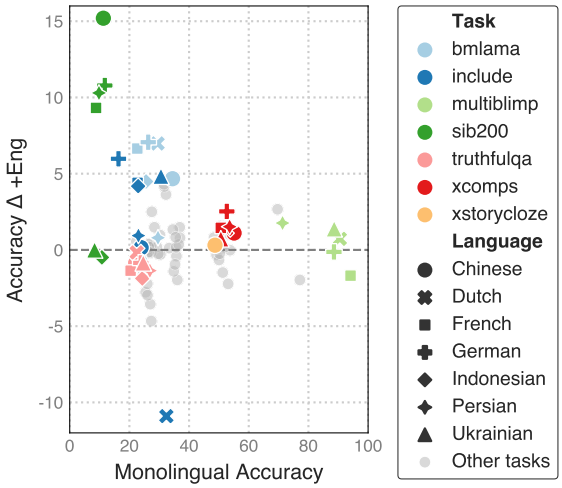

Figure 3: Impact of training LMs on bilingual corpora (adding English) across our evaluation suite. The y-axis denotes the change in accuracy from monolingual to bilingual performance. Dutch SIB-200 performance is omitted due to space constraints (+24.8).

second training language leads to consistent performance gains across most languages. A notable exception is Dutch on INCLUDE, where bilingual training slightly reduces performance. This may be due to domain mismatch: the Dutch corpus includes high-school exam texts, and the addition of English data likely shifts the model away from this domain. Performance on formal linguistic tasks such as MultiBLiMP remains largely unchanged, suggesting that syntactic competence is less sensitive to bilingual input in this setup.

6 Future Outlook

BabyBabelLM is shared research. Starting as a grassroots initiative and conducted in an open and inclusive manner, this resource was gathered by multiple experts with a shared goal. Therefore, we call for this collaboration to continue and welcome further contributions for BabyBabelLM, even after the paper is published. In summary, we hope that BabyBabelLM will serve as a valuable resource for the community, facilitating reproducible, comparable, and cost-effective exploration.

To act as a complement to the resource, we provide a list of potential questions we believe this resource may aid in answering: Do LMs acquire language more like language learners of a specific language than another? Are there critical times for learning a second language in LMs (Constantinescu et al., 2025)? Can we replicate results in studies on the border of linguistics and LLMs, where testing on only a single language might bias results (Arnett et al., 2025)? Is there a way to overcome different scripts and unshared tokenizers and provide the same cross-lingual benefits between languages regardless of differences in script? What is the right tokenization scheme across languages, and is tokenization needed at all (Hwang et al., 2025; Rust et al., 2023)? While humans typically give consistent answers across languages, current LMs often do not (Qi et al., 2023; Goldman et al., 2025). Even when outputs align, internal changes tend to affect only one language, indicating a degree of separation not seen in human cognition (Ifergan et al., 2024). Can that be changed?

We hope that BabyBabelLM will serve as a foundation for addressing the questions outlined above, and we invite the community to build on this resource to advance a more inclusive and systematic understanding of multilingual language acquisition and modeling.

Limitations

Our resources target a diverse array of audiences, and therefore our decisions are bound to not satisfy each of those perfectly. While deciding between practical constraints, data availability, and potential research needs we prioritized what we believed would make research and experimentation in the BabyLM paradigm easier. Still, we view our dataset only as a starting point. There are many more languages to be included, and even for the featured languages we imagine further untapped sources of developmentally plausible data.

Despite our language coverage being broader than usual in NLP (cf. Joshi et al., 2020), many languages—particularly those with limited digital presence—remain underrepresented. Especially lacking are languages common in African countries and those with smaller speaker populations, which, despite our efforts, are still underrepresented in our collection. We provide instructions for submitting new languages in our GitHub and welcome community contributions.

Although we aimed to collect as much cognitively plausible data as possible, we also want to stress that our datasets do not contain the actual language a single native speaker of any of the included languages is exposed to. While our data approximates this input much better than standard pretraining resources (e.g., Wikipedia dumps or datasets like Dolma, Soldaini et al., 2024), the distribution of topics and formats remains only a gross approximation of the diversity experienced by a native learner.

While we calibrate dataset sizes using byte-adjusted thresholds to ensure comparability, the actual composition of developmentally plausible content varies substantially across languages. In several cases, high-quality child-directed speech (CDS) or educational material is unavailable, and we rely more heavily on fallback sources such as subtitles or Wikipedia. This variability may introduce confounds in cross-linguistic analyses and limits the strength of direct typological comparisons. We recommend that future work interpreting model differences across languages take these compositional disparities into account.

Our final limitation is the lack of cross-linguistically available evaluation resources. Many languages are only evaluated on monolingual datasets explicitly created for them, and beyond MultiBLiMP there is currently no resource that covers all included languages. As the study of bilingualism or the acquisition of multiple languages (by models and/or humans) are intended applications of this dataset, we are also constrained by a lack of resources that explicitly target multilingual capabilities. We did not create a standardized testbed to test such questions ourselves, as they are too varied. However, we hope that our data and existing evaluations can serve as inspiration for further research in that direction.

References

2019. Winogrande: An adversarial winograd schema challenge at scale.

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A leveled reading corpus of Modern Standard Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Latifa Al-Sulaiti, Noorhan Abbas, Claire Brierley, Eric Atwell, and Ayman Alghamdi. 2016. Compilation of an Arabic children's corpus. In *Proceedings of the Tenth International Conference on Language*

Resources and Evaluation (LREC'16), pages 1808–1812, Portorož, Slovenia. European Language Resources Association (ELRA).

Iolanda Alfano, Francesco Cutugno, Aurelio De Rosa, Claudio Iacobini, Renata Savy, Maria Voghera, and 1 others. 2014. Volip: a corpus of spoken italian and a virtuous example of reuse of linguistic resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3897–3901. European Language Resources Association (ELRA).

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.

Catherine Arnett, Tyler A Chang, James A Michaelov, and Benjamin K Bergen. 2025. On the acquisition of shared grammatical representations in bilingual language models. *arXiv preprint arXiv:2503.03962*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. Turblimp: A turkish benchmark of linguistic minimal pairs. *Preprint*, arXiv:2506.13487.

Yoshua Bengio, Sören Mindermann, and Daniel Privitera. 2025. International ai safety report 2025.

Ryan Bishop. 1991. There's Nothing Natural About Natural Conversation: A Look at Dialogue in Fiction and Drama. *Oral Tradition*, pages 58–78.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

Bastian Bunzeck and Holger Diessel. 2025. The richness of the stimulus: Constructional variation and development in child-directed speech. *First Language*, 45(2):152–176.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in German BabyLMs? In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 169–186, Vienna, Austria. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, and 29 others. 2023a. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science*, 27(6):843–873.

Thea Cameron-Faulkner and Claire Noble. 2013. A comparison of book text and Child Directed Speech. *First Language*, 33(3):268–279.

Luca Capone, Alice Suozzi, Gianluca Lebani, and Alessandro Lenci. 2024. BaBIEs: A benchmark for the linguistic evaluation of Italian baby language models. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 157–170, Pisa, Italy. CEUR Workshop Proceedings.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *CoRR*, abs/2408.10441.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every Layer Counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 210–224, Singapore. Association for Computational Linguistics.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Jiali Cheng and Hadi Amiri. 2024. Mu-bench: A multi-task multimodal benchmark for machine unlearning. *Preprint*, arXiv:2406.14796.

Madalina Chitez, Mihai Dascalu, Aura Cristina Udrea, Cosmin Strilețchi, Karla Csürös, Roxana Rogobete, and Alexandru Oraviţan. 2024. Towards building the LEMI readability platform for children's literature in the Romanian language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16450–16456, Torino, Italia. ELRA and ICCL.

Leshem Choshen, Yang Zhang, and Jacob Andreas. 2024. A Hitchhiker's Guide to Scaling Law Estimation. *arXiv preprint*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for Chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786, Osaka, Japan. The COLING 2016 Organizing Committee.

Yiming Cui, Ting Liu, Ziqing Yang, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. 2020. A sentence cloze dataset for chinese machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*.

G. William Domhoff and Adam Schneider. 2008. Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition*, 17(4):1238–1247.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *Preprint*, arXiv:2305.07759.

Steven Y. Feng, Noah Goodman, and Michael Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and PyPDF2 Contributors. 2022. The PyPDF2 library.

Achille Fusco, Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi, and Cristiano Chesi. 2024. Recurrent networks are (linguistically) better? an (ongoing) experiment on small-LM training on child-directed speech in Italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 382–389, Pisa, Italy. CEUR Workshop Proceedings.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Volker Gast, Christian Wehmeier, and Dirk Vanderbeke. 2023. A Register-Based Study of Interior Monologue in James Joyce's Ulysses. *Literature*, 3(1):42–65.

Giuliana Genovese, Maria Spinelli, Leonor J. Romero Lauro, Tiziana Aureli, Giulia Castelletti, and Mirco Fasolo. 2020. Infant-directed speech as a simplified but not simple register: A longitudinal study of lexical and syntactic features. *Journal of Child Language*, 47(1):22–44.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, and 1 others. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *arXiv preprint arXiv:2502.21228*.

Zebulon Goriely and Paula Buttery. 2025. IPA CHILDES & G2P+: Feature-rich resources for cross-lingual phonology and phonemic language modeling. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 502–521, Vienna, Austria. Association for Computational Linguistics.

Anna Elizabeth Gowenlock, Courtenay Norbury, and Jennifer M. Rodd. 2024. Exposure to Language in Video and its Impact on Linguistic Development in Children Aged 3–11: A Scoping Review. *Journal of Cognition*, 7(1):57.

François Grosjean. 1989. Neurolinguists, beware! the bilingual is not two monolinguals in one person. *Brain and language*, 36(1):3–15.

Andreas Hallberg. 2025. An 81-million-word multi-genre corpus of arabic books. *Data in Brief*, 60:111456.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Adrian Florea, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, Hinrich Schütze, and Nima Mesgarani. 2025. Xcomps: A multilingual benchmark of conceptual minimal pairs. *Preprint*, arXiv:2502.19737.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. 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*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Julie Hunter, Jérôme Louradour, Virgile Rennard, Ismaïl Harrando, Guokan Shang, and Jean-Pierre Lorré. 2023. The claire french dialogue dataset. *Preprint*, arXiv:2311.16840.

Sukjun Hwang, Brandon Wang, and Albert Gu. 2025. Dynamic chunking for end-to-end hierarchical sequence modeling. *arXiv preprint arXiv:2507.07955*.

Maxim Ifergan, Omri Abend, Renana Keydar, and Amit Pinchevski. 2024. Identifying narrative patterns and outliers in holocaust testimonies using topic modeling. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 44–52, Torino, Italia. ELRA and ICCL.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2024. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Da Ju, Hagen Blix, and Adina Williams. 2025. Domain regeneration: How well do LLMs match syntactic properties of text domains? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2367–2388, Vienna, Austria. Association for Computational Linguistics.

Andreas H Jucker. 2021. Features of orality in the language of fiction: A corpus-based investigation. *Language and Literature: International Journal of Stylistics*, 30(4):341–360.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language

identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Krisjanis Karins, Robert MacIntyre, Monika Brandmair, Susanne Lauscher, and Cynthia McLemore. 1997. CALLHOME German Transcripts.

Daria Kryvosheieva and Roger Levy. 2025. Controlled evaluation of syntactic knowledge in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

Richard Lastrucci, Jenalea Rajab, Matimba Shingange, Daniel Njini, and Vukosi Marivate. 2023. Preparing the vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 18–25, Dubrovnik, Croatia. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022b. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Jing Liu and Abdellah Fourtassi. 2025. Benchmarking llms for mimicking child-caregiver language in interaction. In *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue – Full Papers*, pages 92–103, Bielefeld, Germany. SEMDIAL.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Zhi Liu, Dong Li, Taotao Long, Chaodong Wen, Xian Peng, and Jiaxin Guo. 2025. CSQ: A Chinese Elementary Science Question Dataset with Rich Discipline Properties in Adaptive Problem-Solving Process Generation.

Emiddia Longobardi, Clelia Rossi-Arnaud, Pietro Spataro, Diane L Putnick, and Marc H Bornstein. 2015. Children's acquisition of nouns and verbs in italian: contrasting the roles of frequency and positional salience in maternal language. *Journal of child language*, 42(1):95–121.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, pages 517–540.

Alexis Matzopoulos, Charl Hendriks, Hishaam Mahomed, and Francois Meyer. 2025. BabyLMs for isiXhosa: Data-efficient language modelling in a low-resource context. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 240–248, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marina Mayor-Rocher, Cristina Pozo, Nina Melero, Gonzalo Martínez, María Grandury, and Pedro Reviriego. 2025. It's the same but not the same: Do llms distinguish spanish varieties? *Preprint*, arXiv:2504.20049.

Cindy McKellar. 2022. Autshumato english-sesotho parallel corpora. SADiLaR Language Resource Repository, License: Creative Commons Attribution 4.0 International.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. 2025. Enhancing multilingual llm pretraining with model-based data selection. *arXiv*.

Francois Meyer and Jan Buys. 2024. Triples-to-isiXhosa (T2X): Addressing the challenges of low-resource agglutinative data-to-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16841–16854, Torino, Italia. ELRA and ICCL.

Ludmila Midrigan Ciochina, Victoria Boyd, Lucila Sanchez-Ortega, Diana Malancea_Malac, Doina Midrigan, and David P. Corina. 2020. Resources in underrepresented languages: Building a representative Romanian corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3291–3296, Marseille, France. European Language Resources Association.

Steven Moran. 2016. The ACQDIV database: Min(d)ing the ambient language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4423–4429, Portorož, Slovenia. European Language Resources Association (ELRA).

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artif. Intell. Rev.*, 57(10):265.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Afrobench: How good are large language models on african languages? *Preprint*, arXiv:2311.07978.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-directed language does not consistently boost syntax learning in language models. *Preprint*, arXiv:2505.23689.

Katerina Papantoniou and Yannis Tzitzikas. 2024. Nlp for the greek language: A longer survey. *Preprint*, arXiv:2408.10962.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Velka Popova. 2020. Childes bulgarian labling corpus.

Laurent Prévot, Sheng-Fu Wang, Jou-An Chi, and Shu-Kai Hsieh. 2024. Extending the BabyLM initiative : Promoting diversity in datasets and metrics through high-quality linguistic corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 147–158, Miami, FL, USA. Association for Computational Linguistics.

Ayu Purwarianti, Dea Adhista, Agung Baptiso, Miftahul Mahfuzh, Yusrina Sabila, Aulia Adila, Samuel Cahyawijaya, and Alham Fikri Aji. 2025. NusaDialogue: Dialogue summarization and generation for underrepresented and extremely low-resource languages. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 82–100, Online. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen,

Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

Dimitris Roussis, Leon Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2025. Krikri: Advancing open large language models for greek. *Preprint*, arXiv:2505.13772.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 174–188, Miami, FL, USA. Association for Computational Linguistics.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 2 others. 2024. An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*.

Zhewen Shen, Aditya Joshi, and Ruey-Cheng Chen. 2024. BAMBINO-LM: (bilingual-)human-inspired continual pre-training of BabyLM. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–7, Bangkok, Thailand. Association for Computational Linguistics.

Maria Shvedova and Arsenii Lukashevskyi. 2024. Plug: Corpus of old ukrainian texts. Available at `https://github.com/Dandelliony/pluperfect_grac`.

Johannes Sibeko and Menno Zaanen. 2023. A data set of final year high school examination texts of south african home and first additional language subjects. *Journal of Open Humanities Data*, 9.

Mariana O Silva, Clarisse Scofield, and Mirella M Moro. 2021. Pportal: Public domain portuguese-language literature dataset. In *Dataset Showcase Workshop (DSW)*, pages 77–88. SBC.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Dan Isaac Slobin. 2014. *The crosslinguistic study of language acquisition: Volume 5: Expanding the contexts*. Psychology Press.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Catherine E. Snow and Charles A. Ferguson, editors. 1977. *Talking to Children: Language Input and Acquisition*. Cambridge University Press, Cambridge, MA.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *Preprint*, arXiv:2402.00159.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Maria Spinelli, Chiara Suttora, Adrian Garcia-Sierra, Fabia Franco, Francesca Lionetti, and Mirco Fasolo. 2023. Editorial: Are there different types of child-directed speech? dynamic variations according to individual and contextual factors. *Frontiers in Psychology*, Volume 13 - 2022.

Sabine Stoll. 2020. *Sampling linguistic diversity to understand language development*, pages 247–262. John Benjamins Publishing Company.

Simon Strübbe, Irina Sidorenko, and Renée Lampe. 2025. Comparison of grammar characteristics of human-written corpora and machine-generated texts using a novel rule-based parser. *Information*, 16(4).

Alice Suozzi, Luca Capone, Gianluca E Lebani, and Alessandro Lenci. 2025. Bambi: Developing baby language models for italian. *arXiv preprint arXiv:2503.09481*.

Shira Tal, Eitan Grossman, and Inbal Arnon. 2024. Infant-directed speech becomes less redundant as infants grow: Implications for language learning. *Cognition*, 249:105817.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Agnes Tellings, Micha Hulsbosch, Anne Vermeer, and Antal van den Bosch. 2014. Basilex: an 11.5 million words corpus of dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024. Meltemi: The first open large language model for greek. *Preprint*, arXiv:2407.20743.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14006–14014.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. 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*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Anna Whittle and Elena Nuzzo. 2015. L'insegnamento della grammatica nella classe multilingue. un esperimento di focus on form nella scuola primaria. *Italiano LinguaDue*, 7(1):369–370.

Wikimedia Foundation. 2025. Wikisource. English Wikisource. Accessed July 24, 2025.

Ethan Gotlieb Wilcox, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. 2024. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: Fine-grained corpus for Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918. Association for Computational Linguistics.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Weihao You, Pengcheng Wang, Changlong Li, Zhilong Ji, and Jinfeng Bai. 2024. Ck12: A rounded k12 knowledge graph based benchmark for chinese holistic cognition evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19431–19439.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *CoRR*.

Dongjie Zhou and Tianqing Zheng. 2024. Measurement method research of chinese texts' difficulty based on two-characters continuations. *Plos one*, 19(9):e0309717.

Jiaming Zhou, Shiyao Wang, Shiwan Zhao, Jiabei He, Haoqin Sun, Hui Wang, Cheng Liu, Aobo Kong, Yujie Guo, and Yong Qin. 2024. Childmandarin: A comprehensive mandarin speech dataset for young children aged 3-5. *CoRR*.

A Format considerations

In order to make our multilingual dataset easy to use and access for researchers, we format our data in a unified schema across languages consisting of self-contained documents and document-level metadata. This is applied and verified consistently for all data, including language-specific resources, multilingual datasets, and various corpora used for padding. For each document, we include details about the license and the source of the data, ensuring proper creator attribution and compliance with data sharing licenses. Other fields encode information about a document's content, such as the text's script, target age estimate, content category, and number of tokens. The full schema of our documents and a detailed description of each field, are presented in Table 3.

B Language-specific Details

B.1 Arabic

Dataset Description. In recent years, there has been substantial effort towards advancing NLP for the Arabic languages. However, child-oriented resources and developmentally plausible corpora are still lacking. Some efforts have been made such as "a Compilation of an Arabic Children's Corpus" (Al-Sulaiti et al., 2016), and a "leveled reading corpus of Modern Standard Arabic" (Al Khalil et al., 2018). However, the data has not been made publicly available. Additionally, natural conversation and speech data, which is part of the child's linguistic environment, is often unreleased or available only under a fee. Despite these restrictions, we present a developmentally plausible dataset for Arabic, consisting of children's books and stories, song lyrics, natural conversations, and articles from child wikis. We provide details about each data category in the paragraphs below.

Our Arabic dataset includes a large collection of different language varieties, specifically: Sudan, Egyptian, Yemeni, Meghribi, Iraqi, Levantine, Gulf and written language in Modern Standard Arabic (MSA). Even though spoken Arabic can vary substantially across different regions, with speech being often mutually unintelligible, due to the scarcity of developmentally plausible data we opted to combine all the different dialects into one dataset. An important goal for the next iteration of BabyBabelLM is to release single dialect developmentally plausible datasets for Arabic, incorporating data from recent efforts such as Atlas-Chat (Shang et al., 2025). This will give us the opportunity to study the unique nature of dialects in the Arabic language, and how they interact in terms of language model training and performance in a developmentally plausible setting.

Books, Wiki, News. For books, we include the recently released Arabic Book Corpus (Hallberg, 2025), keeping only the "children stories" category, containing both translated and original titles, mostly from the 20th century. We also include children stories from GlotStoryBooks and Ririro and Children's Wiki articles.

Transcription. Given that songs form a common linguistic input for children, we incorporate in our data the Habibi Corpus (El-Haj, 2020), consisting of song lyrics in a variety of dialects. Additionally, adult speech conversation data was collected from MagicHub[5] for the Yemeni [6] and the Egyptian [7] dialects, as dialogue also contributes in the child's learning environment. Lastly, we include all child directed speech present in CHILDES.

B.2 Bulgarian

Dataset Description. The Bulgarian dataset is a compilation of children's literature accessed via a public library website: `https://chitanka.info`. The Bulgarian BabyLM corpus is the first large-scale corpus of child-appropriate Bulgarian text.

Books, Wiki, News. The `Chitanka` portion of the Bulgarian BabyLM corpus consists of 28M to-

[5] `https://magichub.com/datasets/`

[6] Available here: `https://magichub.com/datasets/yemeni-arabic-conversational-speech-corpus/`

[7] Available here: `https://magichub.com/datasets/egyptian-arabic-conversational-speech-corpus/`

kens, excluding punctuation. To our knowledge, the only other similarly sourced dataset is from CHILDES, which is also included in the Bulgarian BabyLM Corpus. The Chitanka library consists of several categories of books, ranging from science to literature, and has a curated section of Children and Young Adult's literature that the site owners has confirmed are free to distribute.[8] Chitanka's Children and Young Adult Literature sections consists of 670 texts, comprised of novels and short stories (377 texts) poems and riddles (40 texts); fairy tales (169 texts); and other children stories (25 texts); and miscellaneous children and young adult literature (68 texts).[9]

Preprocessing. The individual texts have been cleaned of front and back matter. Each text is provided alongside the link it was scraped from. The largest version of the dataset includes children's literature for various ages, but if one would like to restrict the dataset to a subset of earlier-age appropriate literature, this can be done by restricting the URLs which correspond to the Bulgarian Ministry of Education's programme for second and third grade summer reading (ages 6-8 years), for which the corresponding URLs are listed in the README of the dataset. A final notable detail of the dataset is that the texts are in the Cyrillic alphabet, which should be considered during preprocessing.

Transcription. The Bulgarian portion of CHILDES consists of 94K tokens of Child-Directed Speech (CDS) collected by Popova (2020) for 5 children aged 1-2 years.

B.3 Cantonese

Dataset Description. We compile our Cantonese text corpus by consolidating four publicly available resources: the Hambaanglaang project, the GlotStory Book project, and two Cantonese datasets from CHILDES (HKU-70 Corpus and Lee/Wong/Leung Corpus).

Books, Wiki, News. Hambaanglaang[10] is an open-source repository of Cantonese graded readers created by volunteers. It offers a collection of stories designed for children across five proficiency levels, aiming to support Cantonese literacy and reading skills within the community. Detailed information about this project can be found in its official documentation. GlotStory Book (Hong Kong edition)[11] is a free, open-source literacy site that localizes 40 children's stories—originally from the African Storybook Project—into multiple languages used in Hong Kong's "two scripts / three languages" environment (spoken & written Cantonese, Mandarin, English). Here we only extracted Cantonese. Each story is tagged with one of five length/lexical-complexity levels and accompanied by narrated audio recordings intended to support family-, school-, and community-based language learning. The HKU-70 Corpus[12] contains 70 transcripts of interviews with 70 Cantonese-speaking children aged 2 years 6 months to 5 years 6 months. The data were collected at the University of Hong Kong and represent naturalistic child–adult interactions in preschool settings. Each child participated in a one-hour recording session, with conversations organized around familiar daily routines (e.g., bathing, dressing, feeding) to elicit a diverse range of utterances and syntactic structures. The sample was balanced by gender, and all children were prescreened using the Cantonese version of the Reynell Developmental Language Scales. Finally, the Lee/Wong/Leung Corpus[13], which provides longitudinal data on eight Cantonese-speaking children, each recorded for approximately one year. The recordings capture natural interactions between children, their caregivers, and occasionally other adults. Detailed metadata about the children, including their ages and family backgrounds, are included, providing valuable sociolinguistic context for the dataset.

Preprocessing. All datasets were cleaned to retain only complete Traditional Chinese text, with non-textual annotations such as speaker labels and syntactic tags removed. Each storybook page or conversational block is treated as a single passage-level entry. All text is tokenized using the Qwen1.5-7B (Bai et al., 2023) tokenizer to ensure compatibility with downstream language modeling tasks.

[8] Excerpt from author correspondence: "*Everything in my library is completely free; you're welcome to use any of the available resources. The books we add are supposed to be free of copyright claims. If such claims do arise—that is, if rights holders or distributors get in touch with us—those works are 'quarantined' until a previously agreed period of time has elapsed.*"

[9] Available here: https://chitanka.info/books/category/detska-literatura

[10] https://hambaanglaang.hk/about-us-2/

[11] https://global-asp.github.io/storybooks-hongkong

[12] https://talkbank.org/childes/access/Chinese/Cantonese/HKU.html

[13] https://talkbank.org/childes/access/Chinese/Cantonese/LeeWongLeung.html

B.4 Mandarin Chinese

Dataset Description. In addition to multilingual resources (CHILDES and GlotStoryBook), our Mandarin Chinese dataset includes data from multiple resources.

Books, Wiki, News. We use children's book and stories data collected from various sources. We first obtained children's stories from `Quangushi (full stories)`[14]. Then, we used children's stories from and two Chinese reading comprehension datasets: `CFT` (Cui et al., 2016) and `CMRC-2019` (Cui et al., 2020). These two datasets are respectively Cloze and sentence ordering benchmarks derived from children's stories. We reconstructed the complete stories using the answers provided by the authors and included the stories in our dataset. We also collected open-source children's book and children's wiki data from `WikiJunior` and `Wikibooks`.

Education. For educational materials, we used several datasets that evaluate models' general knowledge through exam-style questions, as these datasets are typically well-documented and come with openly available licenses:

- `GAOKAO` (Zhang et al., 2023): an evaluation framework that uses Chinese National College Entrance Examination (GAOKAO) questions as a dataset to evaluate LLMs. The dataset includes subjective and objective questions from exams from 2010 to 2024.
- `CK-12` (You et al., 2024): an evaluation for Chinese LLMs. Constructed based on multi-level knowledge graph and covers most comprehensive knowledge points in Chinese K12 field.
- `CSQ` (Liu et al., 2025): a Chinese Science Question dataset covering four subjects and multiple topics at the Chinese primary school.

We included the full question prompts, answer choices, correct answers, and explanations. Questions in English are excluded. In addition, we collected grammatical and corrected sentences from `FCGEC` (Xu et al., 2022), a human-annotated corpus based on multi-choice grammatical error problems. We also collected data from a hierarchical corpus of primary school students' compositions (Zhou and Zheng, 2024). The primary source of this corpus is elementary school student composition magazines, which ensures that the essays are of relatively high quality. We also incorporated transcriptions in `ChildMandarin` (Zhou et al., 2024). This dataset contains high-quality speech data collected from 397 children in China, along with carefully crafted, character-level manual transcriptions. We included data from `NaturalConv` (Wang et al., 2021), a multi-turn topic-driven conversation dataset is also included as such conversations on daily topic are considered as children available.

Transcription Finally, we included data from Wenetspeech, a multi-domain Mandarin corpus consisting of transcribed speech, collected from Youtube and Podcast. We only used the high-quality labeled part of the dataset, which shall be considered as children available.

B.5 Dutch

Dataset Description. The Dutch data is built from various educational sources. Licensing laws are very strictly defined in the Netherlands, which makes it challenging to find children's literature with creative commons licenses. Educational resources, however, are often released under CC-BY license.

Books, Wiki, News. We include the texts of all high school exams[15] from 1999 to 2024, for all Dutch high school levels: `VMBO` (age 15–16), HAVO (age 16–17), and `VWO` (age 17–18), resulting in 6.87M tokens. We extracted 8.78M tokens from `WikiWijs`[16], a platform for sharing educational materials by teachers for both primary and high school level. `KlasCement`[17] provides a similar platform, focused on Flemish education, from which we extract 0.14M tokens. Next to these educational resources, we also incorporate `BasiLex (Tellings et al., 2014)` into the Dutch section. `BasiLex` contains a collection of child-directed resources, extracted from children's media, children's books, and educational materials. We collect 11.37M tokens from `BasiLex`.

B.6 French

Dataset Description. In addition to child-directed speech from CHILDES (around 4 million tokens), we include the following developmentally plausible resources, covering a range of spoken and written language that children are likely to hear or read.

Books, Wiki, News. We included eighteen children's books (around 1 million tokens).[18] These

[14] `http://quangushi.com/`

[15] Released publicly by examenblad.nl, with archives available at alleexamens.nl.

[16] wikiwijs.nl

[17] klascement.net

[18] Hand-picked from the Wikisource category: Caté-

were selected to match the reading level of children aged 6 to 12 and to cover a variety of story types. The collection includes classic fairy tales (Contes de Perrault, Grimm, Andersen), simple educational texts (Abécédaire du petit naturaliste, Histoires comme ça pour les petits), and famous adventure and fantasy stories like Le Tour du monde en quatre-vingts jours, L'Île au trésor, Alice au pays des merveilles, and Croc-Blanc. Our data also includes subtitles (around 6 million tokens). This portion is made mostly of subtitles from the popular animated series Caillou, aimed at toddlers and shared on YouTube by the channel `Caillou Français - WildBrain`.[19] It includes 1,539 video episodes. In addition to Caillou, we included other well-known children's shows in France, such as Olive et Tom (171 videos), Lou (52 videos), La vie (8 videos), and a few other youth-oriented clips (15 videos). Each subtitle document includes the YouTube video ID so the original video can be accessed. We obtained raw transcripts via the `YouTubeTranscriptApi`,[20] filtered for manually entered transcripts (as opposed to automatically created ones), and fed them through the library's built-in `TextFormatter`, which strips out all timing information and reassembles each subtitle fragment as plain text. Additionally, we included transcripts of spoken conversations (around 2 million tokens) that are not directly addressed to children, but that children could realistically overhear. We selected a number of sources from `Claire-Dialogue-French-0.1`[21] (Hunter et al., 2023), including three types of settings: spontaneous everyday conversations (in homes, cafés, or on the street), guided one-on-one interviews, and workplace meetings. The data comes from sources like `PFC_free`, `OFROM`, `CLAPI`, `ORFEO_coralrom`, `ParisStories`, `CFPP`, `ACSYNT`, and `ORFEO_reunions_de_travail`.

B.7 German

Dataset Description. Our German data builds upon the existing German BabyLM corpus by Bunzeck et al. (2025), extending it with more developmentally plausible data and discarding the majority of their padding data in favor of the multilingual padding data compiled in the current project. As German is a comparatively high-resource language, we are able to supplement the multilingual resources with a variety of monolingual corpora.

Books, Wiki, News. Five different children's wikis are available for German, including the state-sponsored `Klexikon` and `MiniKlexikon`, but also comparable efforts for Austrian German like the `Kiwithek`. In addition, we supplement this kind of educational data with the `WikiBooks Wikijunior` bookshelf, which is fairly comprehensive for German. As for books, we aim to make an educated selection of the `Project Gutenberg` collection featuring works for children and young adults. We include books that are considered classics of children's literature and read to this day. We further also include classics of German literature that are regularly read in middle school (e.g. works by Franz Kafka). Although they are located at the end of the 'developmentally plausible' timeline, they are plausibly encountered by many young adults in the German education system. Similarly, we also include the archives of the `Fluter` magazine, published by the German Federal Agency for Civic Education, which contains a large body of non-fiction writing aimed at adolescents and young adults. Moving from child-directed to child-available language, we furthermore incorporate the German section of the `CallHome corpus` (Karins et al., 1997), which contains transcribed telephone conversations between adults. Such conversations could i) be plausibly overheard by children, and ii) approximate child-directed input nicely by being transcribed from spoken data, which differs quite dramatically from written data in composition (cf. Cameron-Faulkner et al., 2003; Cameron-Faulkner and Noble, 2013; Bunzeck and Diessel, 2025). In a similar vein, we also incorporate the German portion of `Dreambank` (Domhoff and Schneider, 2008), a large corpus of dream reports by adults and children. Despite not being originally spoken, the 'self-reporting' register included in this data is closer to spoken data than ordinary writing, and social storytelling is an important component of language acquisition. Therefore, we conclude that this dataset also enhances the variety and developmental plausibility of the German data.

B.8 Greek

Dataset Description. NLP for the Greek language has developed drastically over the past few years (Papantoniou and Tzitzikas, 2024), with a notable example being the recent release of large language

gorie:Littérature jeunesse

[19] https://www.youtube.com/@CaillouFrench

[20] https://pypi.org/project/youtube-transcript-api/

[21] https://huggingface.co/datasets/OpenLLM-France/Claire-Dialogue-French-0.1

models for the Greek language: Meltemi 7B (Voukoutis et al., 2024) the first such open LLM for Greek, and Krikri 8B (Roussis et al., 2025) further scaling up data and model sizes. Here we present, to our knowledge, the first developmentally plausible corpus for the Greek language. The data is curated as a collection of publicly available datasets, sourced mostly from CLARIN:EL[22], and original web-scraped children's books and stories. We present details about the dataset composition and preprocessing below. In the future we plan to include more child-directed speech data in collaboration with language acquisition researchers, incorporating efforts such as the Greek Children Spoken Language Corpus [23] and the Greek-speaking Children Corpus [24].

Education. We incorporate into our data a variety of educational textbooks. We include a selection of Primary School Books [25] in the fields of arts, language, religion, history, and social and political sciences, aimed at grades 1-6 (ages 6-12). Apart from textbooks aimed at children, we decided to additionally include material designed for later grades and ages. Even though this content is aimed at the tail end of our target ages for developmentally plausible corpora, we consider it sufficiently relevant and representative of the linguistic input of children. Thus, we collect the CGL Modern Greek Texts corpus [26], which comprises around 2 million words from textbooks published by the Greek Ministry of Education taught through grades 7-12 (ages 13-18) in the public school system. We also include the corpus of Pedagogical Greek L2 textbooks [27], addressed to indigenous populations or minorities learning Greek as a second language, aimed at proficiency levels A1 to C2 and ages 6-18+. Even though this resource is designed for non-native learners, we believe the material to be sufficiently close in nature to the learning resources for native Greek speakers. Finally, we include articles from Children's Wikis.

Books, Wikis, News. Numerous websites host open access children e-books and children stories for the Greek language. We identified `openbook.gr`[28] and `free-ebooks.gr`[29] as the largest such sites, and manually scraped them, selecting e-books from the categories of `children`, `young-adult`, and `preschool-education`. The data consists of children books in the Public Domain, as well as open access books released with permissive licenses. We also include a collection of children stories scraped from `paidika-paramythia.gr` [30]. The site enables any author to make a submission in collaboration with the moderators, and includes stories from tradition and mythology, as well as original entries. Lastly, we include sort stories provided in the GlotStoryBooks corpus.

Transcription. We collect publicly available data corresponding to child-produced and child-directed speech. Child Speech [31] contains transcriptions of children's speech with a focus on narration; as the result of interviews conducted by university students with children related to them either by friendship or kinship. Our second addition is the Greek Student Chat Dataset [32] consists of chat among students (grades 4-18) in online collaborative learning environments (wikis). Finally, we also include the Greek portion of CHILDES noting that speech is transcribed in the Latin script. In future efforts we plan on either removing this data or transliterating it to the Greek script.

Everyday conversations between adults are a natural stimulus for children during language development. We include in our data a corpus of written transcripts of everyday conversations between students of the Department of Linguistics [33] that took place between 2001 and 2006. The data is further supplemented by the Babiniotis archive [34], consisting of the same data variety recorded in 2020. The speech is authentic and idiomatic with speakers labeled, resulting in a high quality spoken Greek corpus.

Preprocessing. For the education datasets in our corpus, standard pre-processing was applied. Notably, the Primary School Books corpus re-

[22] https://inventory.clarin.gr/
[23] http://gcsl.ece.uth.gr/
[24] https://gavriilidou.gr/greek-speaking-children-corpus/
[25] https://inventory.clarin.gr/corpus/1075
[26] http://hdl.handle.net/11500/KEG-0000-0000-24FD-B
[27] http://hdl.handle.net/11500/ATHENA-0000-0000-2631-E
[28] https://www.openbook.gr/literature
[29] https://free-ebooks.gr/tag/16?
[30] https://www.paidika-paramythia.gr/16
[31] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-610D-5
[32] http://hdl.handle.net/11500/IONION-0000-0000-5E14-1
[33] http://hdl.handle.net/11500/UOA-0000-0000-5D9C-9
[34] http://hdl.handle.net/11500/UOA-0000-0000-2515-F

quired considerable cleaning and normalization efforts, containing web-scraping artifacts such as javascript code. Regarding e-books, processing the text proved challenging, and required a substantial amount of manual labor. Initially licensing information was extracted, and the corpus was filtered to include only permissive licenses (e.g., cc-by-nc). For openbook.gr, license information is provided as metadata for each entry, while for free-ebooks.gr we manually annotate each book with its license as stated in the text. The stories in paidika-paramythia.gr are released as Public Domain. The text is first extracted from e-books using PyMuPDF[35], and is then filtered to remove license statements, author biographies, and other information deemed irrelevant. Further document-specific normalization follows, fixing text extraction errors, removing unwanted unicode characters, and ensuring the validity of the book content. As part of this process, documents deemed unsuitable for children are excluded. As for the transcription data, standard pre-processing was applied. Morphological and other linguistic annotations where removed from speech data. We note that to ensure anonymity, placeholders exist in conversational text that substitute real information (e.g., names, locations).

B.9 Italian

Dataset Description. Interest in developmentally plausible NLP models for Italian has recently increased, as shown by new training setups and evaluation resources targeting child-directed-language (Fusco et al., 2024; Suozzi et al., 2025; Capone et al., 2024). In assembling our corpus beyond the multilingual resources described in the body of the paper, we enrich it with a set of Italian-specific materials.

Books, Wikis, News. We were able to include approximately thirty books from the independent Italian publishing house Biancoenero Edizioni [36], which kindly shared them with us upon request. The publisher has long been committed to the Alta Leggibilità ("High Readability") project, aimed at making books accessible to all children, including those with reading difficulties. All books are written by Italian authors and are targeted at readers between the ages of 4 and 10. The themes span a range of topics including environment and ecology, bullying, mystery, diversity and inclusion, growing up and intergenerational relationships, and adventure, according to the categories listed in the publisher's updated catalog. In addition to these recently published works, we also incorporate books from the Logos Group library [37]. This collection comprises classic children's stories and fairy tales authored by both Italian and foreign writers whose works are translated into Italian. The estimated target reading age for these texts ranges from approximately 6 to 14 years; however, some of these stories may be orally presented to younger children. Furthermore, we include a series of fairy tales (all from copyright expired sources) curated by the researchers in this study (Fusco et al., 2024). The book section concludes with a manually curated selection of approximately 50 titles from the Project Gutenberg catalog. These works are either explicitly included in the national curriculum for lower and upper secondary education, or authored by canonical figures whose writings are frequently excerpted in educational contexts and whose titles are broadly recognized within the Italian school system. These include both Italian and non-Italian authors. Although the language used in these works is occasionally archaic and stylistically distant from contemporary Italian, as similarly observed in the case of German (and other languages, where applicable), their inclusion aligns with the upper boundary of the "developmentally plausible" timeline. Nevertheless, these texts remain realistically encountered by a substantial portion of young adults within the Italian educational system. To complement these literary sources, we also include the Italian portion of the WikiBooks Wikijunior bookshelf. This collection comprises a range of accessible entries on diverse topics (e.g., the human body, dinosaurs, the solar system), thereby extending our coverage of child-oriented reading materials beyond narrative texts.

Education. Our educational resources cover a range of materials reflecting both formal and informal learning contexts. We begin with standardized assessments, including the Italian portion of past INVALSI tests in Italian and Math at both primary and secondary levels[38] . INVALSI is the national body responsible for evaluating student competen-

[35] https://github.com/pymupdf/PyMuPDF
[36] https://www.biancoeneroedizioni.it/
[37] https://children.logoslibrary.eu/
[38] https://www.invalsi.it/invalsi/index.php

cies and the quality of the education system. In addition, we incorporate an archive of national high school final examination prompts released by the Italian Ministry of Education [39] , covering the past 20 years. These standardized exams, taken by all students aged 18–19 to obtain their diploma, vary across school types but collectively represent the curricular exposure of the vast majority of Italian students and offer a representative snapshot of the competencies expected of young adults within the national system. Finally, we leverage a dataset previously curated by Suozzi et al. (2025) that includes children's songs from the Zecchino D'Oro archive, a long-standing and renowned Italian music festival for children. In addition, we include around 60 YouTube video transcripts from the animated cartoon Calimero [40]. Cartoons, while primarily designed for entertainment, also foster and support children's language development. Our selection prioritizes episodes with consistent and realistic punctuation, filtering out automatically generated transcripts containing grammatical errors or typos.

Lastly, we complement these resources with two text simplification datasets. The first, from Brunato et al. (2015), comprises Terence and Teacher: Terence contains 32 short children's stories with expert-produced simplifications for readers with comprehension difficulties, while Teacher includes 18 pairs of simplified and original texts from a variety of genres (e.g., literature, textbooks). The second dataset, MultiLS, was developed for the MLSP2024 shared task (Shardlow et al., 2024) and focuses on lexical simplification.

Transcription – Child Directed Speech. We use transcripts from psycholinguistic studies on child language acquisition (Longobardi et al., 2015; Whittle and Nuzzo, 2015; Spinelli et al., 2023), which have already been employed as training data in Suozzi et al. (2025). These materials originate from in vivo conversations recorded during experimental sessions and consist of utterances produced by caregivers and directed to children.

Transcription – Child Available Speech. In addition to direct caregiver input, we also consider speech that children are indirectly and routinely exposed to in their environment by overhearing adult–adult conversations. To capture this dimension, we incorporate an open-source dataset comprising 10.43 hours of transcribed conversational speech on specific topics[41] , as well as VolIP, a dataset of telephone conversations (Alfano et al., 2014) already used in Fusco et al. (2024) work.

B.10 Japanese

Dataset Description. In addition to multilingual resources, our Japanese dataset includes educational content from Wikibooks[42] and Wikijunior[43], as well as children's books from Aozora Bunko[44].

Books, Wiki, News. For educational materials, we took data from `Wikibooks`. We used the "Elementary School Learning" section, which targets Japanese elementary school students, typically aged 6 to 12. The content covers major school subjects, including the Japanese language, social studies, mathematics, and science. We excluded pages that were still under construction, as well as those consisting primarily of numerical content(e.g., math drills). The resulting `Wikibooks` corpus contains approximately 0.2M words. We also used `Wikijunior`, which offers educational content designed for Japanese children aged approximately 8 to 11. As with `Wikibooks`, we excluded pages that were under construction or contained only numerical content. The final `Wikijunior` corpus consists of 75 pages, totaling approximately 0.07M words from `Wikijunior`. We complemented our data with texts from `Aozora Bunko`, a Japanese digital library that provides access to literary works in the public domain. We used the aozorabunko-clean dataset[45], a cleaned version of the original collection, that includes only books whose copyrights have expired. This dataset contains storybooks, biographies, poetry, and other literary genres, with the majority being storybooks. It also contains Japanese translations of foreign literature. From this dataset, we selected only children's books. The list of children's book titles was scraped from the category-wise list of titles on `Aozora Bunko`[46]. Books written in old character forms were excluded. This subset comprises 1,111 titles and totals approximately 8.7M words.

[39] https://www.mim.gov.it/
[40] https://www.youtube.com/@calimeroitaliano2815
[41] https://magichub.com/datasets/
[42] https://ja.wikibooks.org/wiki/
[43] https://ja.wikibooks.org/wiki/Wikijunior
[44] https://www.aozora.gr.jp/
[45] https://huggingface.co/datasets/globis-university/aozorabunko-clean
[46] https://yozora.main.jp/

B.11 Persian

Dataset Description. Our Persian dataset includes several curated subcategories designed to support both child-centered and educational language modeling. The final collection contains about 98.5 million words across 217,880 records and consists of four parts: Children's Books, Educational Documents, Child-Directed Speech, and Subtitles used as supplementary padding.

Books, Wiki, News. To construct a subset of educational documents, we started with `FineWeb2-HQ` (Messmer et al., 2025), a high-quality, multilingual dataset built on top of `FineWeb2` (Penedo et al., 2025), as the base for our educational subset. To identify educational content within the Persian subset, we fine-tuned an XLM-R (Conneau et al., 2020) model using a regression task inspired by the `FineWeb-edu` (Lozhkov et al., 2024) methodology. The training data for this model were annotated using Qwen2.5-72B-Instruct (Qwen et al., 2025), following a 5-point additive rubric designed to assess the educational suitability of a document for primary to grade school learners. The documents were then scored between 0 and 5, which were later normalized to the 0–1 range. We trained the XLM-R model to predict these normalized scores. For our final dataset selection, we applied the trained model to `Persian FineWeb2-HQ` documents. We selected documents that (1) were under 3,000 words in length (to avoid structural drift across sections), (2) had a quality score of at least 0.35 based on `FineWeb2-HQ` metadata, and (3) received a predicted educational score of 0.9 or higher. This filtering ensures that the selected documents meet at least the first four points of the rubric, which means they are coherent, suitable for grade-school learners, and contain well-structured educational material. Our subset of children's books includes child-friendly Persian texts sourced from two main corpora: `Ririro` (Persian section) [47] and `GlotStoryBook`. We scraped texts from the Persian section of the `Ririro` story collection. Although the content was mostly clean, we noticed minor inconsistencies in orthography and annotation. We applied light post-processing to fix punctuation, normalize spelling, and standardize diacritics, resulting in a clean and consistent corpus. `GlotStoryBook` included several very short entries, such as single-word or phrase-level records. To ensure data quality and narrative coherence, we filtered out all records with fewer than three words, resulting in the removal of 123 entries from an original total of 1,150. Notably, about one-third of the `GlotStoryBook` dataset consists of "fa-diacritics" texts, which are Persian sentences written with full diacritics. These fully vocalized texts are typically used in early literacy education in Persian-speaking contexts, particularly during the first stages of primary school. They are the first form of written Persian encountered by children as they begin learning to read and write, offering a bridge toward later reading of undiacritized Persian. Their inclusion enriches the dataset with pedagogically relevant material closely aligned with actual educational practice in early schooling.

For child-directed speech, we utilized Persian transcripts from the `CHILDES` project. Notably, although the spoken language is Persian, the transcripts are written in Latin script using phonetic representations, a transcription style known as Romanized Persian. Apart from standard normalization and deduplication, no further preprocessing was applied to preserve the phonetic and linguistic characteristics of child-directed speech.

To meet our target budget of approximately 100 million words, we supplemented the dataset with Persian subtitle data. Subtitles were selected for their syntactic diversity and colloquial tone, helping to enrich the stylistic and lexical range of the dataset. The subtitles act as neutral padding rather than targeted educational or child-focused content.

B.12 Polish

Dataset Description In addition to the multilingual resources, we add three further data sources to the Polish data. Besides these resources, we were unable to find further child-available data. Unfortunately, no spoken Polish corpora are freely available. Although some larger Polish corpora exist, projects like the National Corpus of Polish only offer rudimentary search functions and no accessible data for our purposes.

Books, Wiki, News. The `Wolne Lektury` archive contains a large number of Polish ebooks. We systematically scraped all virtual bookshelves that contain child-directed/child-available literature and included all ebooks that could be plausibly encountered by children currently learning Polish. In order to do so, we consulted native speakers of Polish and articles on classical Polish children's literature. Furthermore, we opt to include books that are translations of global children's classics

[47] https://ririro.com/fa/

(e.g. *Tom Sawyer* or *Alice's Adventures in Wonderland*), as they could plausibly encountered by children learning Polish. Besides these ebooks we also include all educational materials from the `WikiJunior` bookshelf of Polish `Wikibooks`, and educational materials from the Polish `Wikikids` website, which – despite its name – is not a classical wiki, but rather a general educational website.

B.13 Portuguese

Dataset Description. We present a first iteration of a developmentally plausible dataset for Portuguese. During our initial collection efforts a variety of potentially relevant resources were found, but due to time constraints have not been included in this iteration of the data. Two such resources are PPORTAL, the Public Domain Portuguese-language Literature Dataset (Silva et al., 2021), and a collection of natural speech data from CORAA[48].

Books, News, Wiki. Our BabyLM dataset for the Portuguese language consists primarily of sort stories from GlotStoryBooks and Ririro, and articles from a children's wiki.

Transcript. We include child-directed speech from the Portuguese portion of CHILDES, We supplement this data with conversational spoken Brazilian Portuguese speech [49].

B.14 Romanian

Dataset Description. The Romanian BabyLM corpus consists of texts from CHILDES . We additionally include data from two pre-existing resources. Chitez et al. (2024) introduce the LEMI Romanian children's literature corpus, which consists of 33,154 words. We also include data the children's portion of the Romanian Language Corpus collected by Midrigan Ciochina et al. (2020). [50] The corpus consists of children's literature in its poetry and fairy tales section.

B.15 South African languages: Afrikaans, isiXhosa, isiZulu, Sesotho, Sepedi

Dataset Description. The number of large-scale datasets and benchmarks for African languages has grown in recent years, but the African continent remains under-resourced and under-represented in NLP research (Ojo et al., 2025). Collecting BabyLM datasets for African languages presents several challenges. Besides lacking child-directed speech corpora, most African languages lack even domain-general datasets of sufficient quality and scale to approximate developmentally plausible training.

As a first attempt to create BabyLM datasets for African languages, we focus specifically on the linguistically diverse context of South Africa. South Africa has 12 official languages, some of which are commonly included in massively multilingual web-scraped datasets. Importantly, all languages have some high-quality, manually curated datasets that are publicly available. This is thanks to government initiatives, such as the `South African Centre for Digital Language Resources (SADiLaR)`[51], which prioritise the development of language resources in all official languages. After surveying available datasets across languages, we conclude that five languages are candidates for BabyLM datasets with meaningful amount of data: Afrikaans, isiXhosa, isiZulu, Sesotho (Southern Sotho), and Sepedi (Northern Sotho).

Afrikaans is comparably more resourced and we were able to collect a tier 2 corpus. For the other four languages, we were limited to tier 3 corpora. The proportion of data that is truly developmentally plausible varies between languages and, in some cases, falls short in comparison to higher-resourced languages. While limited in scale, our datasets demonstrate the practical feasibility of BabyLM research for low-resource languages. We hope our work serves as a starting point for future research on developmentally plausible language modelling for African languages.

Books, Wiki, News. Only Afrikaans and Sesotho are represented in `CHILDES`, but we obtain child-directed and child-adult interaction corpora for all five languages from the `SA-CDI` project.[52] We include children's books for all five languages from the `GlotStoryBook` dataset (Kargaran et al., 2023), originally scraped from `Nalibali`[53], an initiative promoting children's literacy in South Africa. For educational content, we include high school exams (Sibeko and Zaanen, 2023) for language subjects (home language and first additional language) for all five languages and `QED` (Abdelali et al., 2014) for Afrikaans, isiXhosa, and isiZulu.

[48] https://sites.google.com/view/tarsila-c4ai/coraa-versions
[49] https://magichub.com/datasets/brazilian-portuguese-conversational-speech-corpus/
[50] https://lmidriganciochina.github.io/romaniancorpus/
[51] https://repo.sadilar.org/
[52] https://sa-cdi.org/
[53] https://nalibali.org/

For isiXhosa, we include the descriptive sentences in T2X (Meyer and Buys, 2024), a data-to-text dataset containing simplified isiXhosa sentences describing (subject, relation, object) triples in a knowledge base.

To match the target dataset sizes (tier 2 Afrikaans and tier 3 for isiXhosa, isiZulu, Sesotho, and Sepedi), we include additional high-quality data to supplement the developmentally plausible data as needed. For Afrikaans, we include OpenSubtitles (Lison and Tiedemann, 2016). For all five languages we include language-specific Wikipedia corpora. This still leaves us short for Sesotho and Sepedi. For Sepedi, we include government news articles from Vukenzele (Lastrucci et al., 2023). Finally, we use sentences from parallel corpora for machine translation to reach our target sizes for Sesotho and Sepedi. For Sepedi we include the highest quality Sepedi sentences in WMT22 (Adelani et al., 2022), as measured by language identification score. For Sesotho we include sentences from the Autshumato English-Sesotho Parallel Corpus (McKellar, 2022).

B.16 Indonesian and its local languages: Javanese, Sundanese, Balinese, Buginese, Makassarese, Minangkabau, Acehnese

Dataset Description. Recent years have seen a significant increase in resources for Indonesian and its local languages, mainly due to collective efforts by NusaCrowd (Cahyawijaya et al., 2023a) and SEACrowd (Lovenia et al., 2024). These initiatives have contributed a wide range of datasets, including conversational corpora, written texts, and multilingual collections. However, developmentally plausible and child-related data are still lacking. Below, we describe the data resources we found.

Transcription. We can only find one dataset available from the aforementioned collective efforts: ASR-INDOCSC, which consists of 4.5 hours of daily conversational speech from children in Indonesia, along with multilingual resources.

Books, Wiki, News. The main sources for cognitively and developmentally plausible data for Indonesian and its local languages come mainly from books obtained from a repository provided by the Ministry of Education & Culture[54]. These are primarily educational books and storybooks for children aged 2 to 12. Since these books are in PDF format, we used PyPDF2 (Fenniak et al., 2022) and Tesseract (Smith, 2007) to extract their content. For data preprocessing, we use Gemma3-27B (Team et al., 2025) for content filtering in three steps: filter out non-child-related books, clean and reformat the extracted book content, and then remove non-child-related content. After cleaning, GlotLID v3 (Kargaran et al., 2023) was used for language detection and grouping, allowing data collection for Javanese, Sundanese, Balinese, Buginese, Makassarese, Minangkabau, and Acehnese. Another major source is the Bobo children's magazine[55], which contains child-targeted articles from January 2020 to May 2025, all of which are exclusively in Indonesian. In addition to these, we incorporated data from multilingual resources, specifically GlotStoryBook (Kargaran et al., 2023) and Ririro[56], for Indonesian language data.

To pad the data and reach the required tiers, OpenSubtitles (Lison and Tiedemann, 2016) data were utilized for Indonesian to reach Tier 1. For local languages to reach Tier 3, we prioritized high-quality, manually curated datasets from NusaX (Winata et al., 2023), NusaWrites (Cahyawijaya et al., 2023b), and NusaDialogue (Purwarianti et al., 2025), followed by Wikipedia and MADLAD-400 (Kudugunta et al., 2023) data for additional padding as needed.

B.17 Spanish

Dataset Description. As the predominant language in 21 countries, Spanish is a pluricentric language and exhibits rich diatopic variations. Far from being a homogeneous language, it encompasses a wide range of national and regional varieties, marked by distinct morphosyntactic and lexical features (Mayor-Rocher et al., 2025). As such, the term "Spanish" does not denote a single standardized form, but rather a set of linguistic norms shaped by diverse cultural and geographic contexts. The resources compiled in this dataset reflect this inherent diversity: our search for developmentally plausible materials was deliberately international, resulting in the inclusion of content from at least eight different countries.

Books, Wiki, News. A substantial portion of children's books is sourced from the Elejandría collection [57], which features 19 translated bed-

[54] https://repositori.kemdikbud.go.id

[55] https://bobo.grid.id

[56] https://ririro.com

[57] https://www.elejandria.com/coleccion/

time stories from classical authors like Andersen, Grimm, and Perrault; 20 translated young adult classics, including "Gulliver's Travels" and "Alice in Wonderland"; and 35 original Spanish-language books by authors from Spain, Uruguay, Mexico, Nicaragua, Cuba, and Argentina, categorized under Discovering Spain and Hispanic American Literature. Additional books were sourced from the Logos Group library, which granted us access upon request. This collection includes Spanish translations of well-known children's literature, such as The Adventures of Tom Sawyer, as well as a smaller number of original Spanish texts. It also features songs, traditional Christmas carols, legends, and famous fables like The Ant and the Grasshopper. Our dataset also includes a range of children's stories, fairy tales, poems, traditional literature, and songs accessed via the Ministries of Education of Argentina [58] and Colombia [59], the provincial government of Salta in Argentina [60], and the educational website *educ.ar.portal* [61].

To capture spoken Spanish that is accessible to children, we incorporated two complementary resources. First, we included an open-source dataset from MagicHub[62], comprising 5.56 hours of transcribed conversational speech in Peninsular Spanish. This dataset features 17 dialogues recorded between four pairs of speakers, covering a variety of everyday topics. Additionally, we incorporated the SpinTX video archive [63], which offers curated video clips and transcripts from the Spanish in Texas Corpus. This collection of interviews with bilingual Spanish speakers residing in Texas covers a wide range of topics relevant to daily life, including family, friendship, food, culture, parenting, education, and school.

B.18 Ukrainian

Dataset Description. The Ukrainian dataset is a collection of different resources. To the best of our knowledge, there is no CHILDES-like corpus for the Ukrainian language; therefore, it has been substituted with a set of monolingual and multilingual data.

Books, Wiki, News. For the majority of developmentally plausible data, we use the GRAC corpus (Shvedova and Lukashevskyi, 2024). This corpus consists of copyright-free texts concerning Ukraine till 1954. The dataset is heavily filtered, reducing from 100M tokens to 29M, to extract the most developmentally plausible data. First, language filtering restricts content to Ukrainian, excluding all other languages, including English, German, Russian, and others. Style-based filtering removes journalistic content, personal memoirs, religious materials, public speeches, official documents, and texts with unknown style classifications. Additionally, non-fiction works published before 1900 are excluded to maintain temporal relevance. The remaining texts are categorized into educational content (academic materials and popular science works), child-appropriate books (fiction, folklore, and poetry), and other materials (internet communication and private oral content). Additionally, we utilize the Ukrainian portion of Wikisource (Wikimedia Foundation, 2025) as a source of fairy tales and fiction books, thereby expanding the dataset by an additional 1 million tokens.

To expand the developmentally plausible data, we incorporate the previously mentioned GlotStorybook and Ririro datasets. Wikipedia serves as a significant source of encyclopedic content, contributing approximately 29.1M tokens. The FineWeb-C corpus provides an additional 174K tokens of contemporary language use. Finally, OpenSubtitles contributes nearly 29.5M tokens of conversational Ukrainian text from movie and television subtitles, to which a child would most likely be exposed.

B.19 Other Languages

For the rest of the languages in the BabyBabelLM dataset, no language-specific resources were collected. Instead, these languages are populated by multilingual data resources, namely: CHILDES, GlotStoryBooks, Ririro, and Child Wikis. These languages are: *Basque, Croatian, Czech, Danish, Estonian, Hebrew, Hungarian, Icelandic, Korean, Norwegian, Romanian, Russian, Serbian, Turkish, Swedish*, and *Welsh* for a total of 16 out of 45 languages. We welcome contributions for these, and other languages, details presented in the project website.

[58] https://www.argentina.gob.ar/educacion/historiasxleer
[59] https://v1.maguared.gov.co/serie-leer-es-mi-cuento-todos-los-titulos/
[60] https://planeamiento.edusalta.gov.ar/
[61] https://www.educ.ar/
[62] https://magichub.com/datasets/
[63] https://spintx.org/

Language	ISO 639-3	Tier	Byte Premium	Actual Data Size (MB)	Transcription Tokens	Education Tokens	Books, Wiki, News Tokens	Subtitles Tokens	Padding Tokens	Total Tokens
Chinese	zho	Tier 1	0.936	518.85	108,368,032	13,037,234	16,429,780	0	0	137,835,046
French	fra	Tier 1	1.174	634.88	6,234,743	0	13,987,611	5,750,144	100,608,287	126,580,785
Bulgarian	bul	Tier 1	1.812	981.07	0	0	24,799,312	0	90,563,381	115,362,693
Indonesian	ind	Tier 1	1.179	638.87	17,824	62,188	17,225,662	0	96,044,750	113,350,424
Dutch	nld	Tier 1	1.052	569.49	3,304,756	19,146,045	17,428,015	971,334	69,035,414	109,885,564
German	deu	Tier 1	1.054	568.98	8,518,785	257,233	7,655,975	0	91,478,846	107,910,839
English	eng	Tier 1	1.0	539.18	36,814,704	0	41,357,314	0	20,706,303	98,878,321
Persian	fas	Tier 1	1.597	867.3	0	94,320,928	67,165	0	4,117,988	98,506,081
Ukrainian	ukr	Tier 1	1.751	945.57	0	12,003,085	16,786,100	0	58,804,062	87,593,247
Japanese	jpn	Tier 2	1.322	71.78	0	291,053	9,712,521	0	6,520,750	16,524,324
Serbian	srp	Tier 2	0.826	77.21	1,489,908	0	29,896	0	13,707,246	15,227,050
Cantonese	yue	Tier 2	0.862	43.34	2,982,684	0	191,861	0	11,870,650	15,045,195
Portuguese	por	Tier 2	1.098	59.5	956,441	0	382,562	0	10,348,599	11,687,602
Swedish	swe	Tier 2	1.021	55.21	750,286	0	526,330	0	9,810,043	11,086,659
Greek	ell	Tier 2	1.967	106.81	1,945,604	2,577,703	1,402,782	0	4,956,467	10,882,556
Polish	pol	Tier 2	1.077	58.29	1,257,155	0	48,831	0	8,906,729	10,212,715
Estonian	est	Tier 2	0.968	52.37	1,026,491	0	0	0	8,814,184	9,840,675
Spanish	spa	Tier 2	1.084	58.75	2,978,384	0	5,385,855	0	1,344,853	9,709,092
Italian	ita	Tier 2	1.067	57.96	1,189,631	990,522	6,797,154	0	490,380	9,467,687
Afrikaans	afr	Tier 2	1.037	56.28	272,434	116,380	153,914	464,009	8,300,684	9,307,421
Welsh	cym	Tier 2	1.027	55.39	1,109,683	0	0	0	7,602,459	8,712,142
Basque	eus	Tier 2	1.06	57.06	201,402	0	1,716,026	0	6,271,869	8,189,297
Korean	kor	Tier 3	1.293	7.07	2,163,779	0	15,458	0	273,838	2,453,075
Sesotho	sot	Tier 3	1.166	6.31	420,926	106,253	141,012	0	557,003	1,225,194
Sepedi	nso	Tier 3	1.116	6.06	0	92,589	122,083	0	871,565	1,086,237
Buginese	bug	Tier 3	1.228	6.67	0	0	41,174	0	961,405	1,002,579
Romanian	ron	Tier 3	1.115	6.1	294,696	0	284,101	0	393,308	972,105
Acehnese	ace	Tier 3	1.242	6.74	0	0	242,613	0	725,581	968,194
Javanese	jav	Tier 3	1.147	6.23	0	0	307,282	0	645,365	952,647
Balinese	ban	Tier 3	1.27	6.87	0	0	63,826	0	874,899	938,725
Icelandic	isl	Tier 3	1.154	6.27	452,099	0	0	0	470,031	922,130
Croatian	hrv	Tier 3	0.99	5.39	469,078	0	0	0	445,976	915,054
Makassarese	mak	Tier 3	1.251	6.79	0	0	34,080	0	873,230	907,310
Nowegian	nor	Tier 3	1.125	6.11	404,670	0	290	0	496,473	901,433
Sundanese	sun	Tier 3	1.097	5.96	0	177	17,264	0	874,647	892,088
Danish	dan	Tier 3	1.021	5.53	372,836	0	8,848	0	457,152	838,836
Hebrew	heb	Tier 3	1.355	7.37	309,854	0	0	0	509,056	818,910
Minangkabau	min	Tier 3	0.95	5.16	0	0	122,536	0	663,669	786,205
Czech	ces	Tier 3	1.036	5.64	377,313	0	0	0	385,263	762,576
isiZulu	zul	Tier 3	1.164	6.31	62,772	56,641	96,383	5,402	532,402	753,600
Russian	rus	Tier 3	1.823	10.0	0	0	92,462	0	655,911	748,373
Hungarian	hun	Tier 3	1.02	5.55	391,041	0	6,234	0	322,636	719,911
isiXhosa	xho	Tier 3	1.199	6.52	74,950	65,208	98,144	29,099	412,004	679,405
Turkish	tur	Tier 3	1.044	5.72	248,397	0	11,193	0	405,478	665,068

Table 2: Detailed data statistics for all languages in the BabyBabelLM dataset. Tiers indicate target size equivalence to English tokens: Tier 1 (100M), Tier 2 (10M), Tier 3 (1M).

C GPT-BERT

We trained monolingual GPT-BERT (Charpentier and Samuel, 2024) models on all our languages in Tier 1 and 2. Models were trained for 500 steps on Tier 1 languages ($\tilde{1}0$ epochs) and for 250 steps on Tier 2 languages ($\tilde{2}5$ epochs). All models had a vocab size of 16,384, 12 layers, 768 hidden size, and 2560 intermediate size.

We report the results in 4, plotting GPT-BERT against the GPT-2 models on SIB-200 and MultiBLiMP. As can be seen, our GPT-2 models consistently outperform GPT-BERT. We leave a more extensive exploration into finding a more optimal GPT-BERT configuration open for future work.

Field	Type	Values	Description
text	string	*Una volta, c'erano...*	The content of the document.
category	string	**Transcription**	
		• child-directed-speech	Speech directed to children and speech produced by children.
		• child-available-speech	Speech children are exposed to without being the target recipients (e.g., adult conversations).
		Education	
		• educational	School textbooks, exams, and other educational material designed for children.
		Wiki, News, Books	
		• child-books	Books and stories created for children.
		• child-wiki	Children wiki articles.
		• child-news	News directed to children.
		Subtitles	
		• subtitles	Subtitles for child-appropriate material (e.g., children TV shows).
		• qed	Subtitles from the QED dataset.
		Padding	
		• padding-wikipedia	Wikipedia articles.
		• padding-[placehorder]	Other forms of padding, used primarily for low-resource languages.
data-source	string	CHILDES, www.ririro.com, ...	The source of the document: dataset name, url, or item identifier.
script	string	Latn, Grek, Cyrl, ...	The script of the text: a validated ISO-15924 code string.
language	string	por, ell, bul, ...	The language of the text: a validated ISO-639-3 code string.
age-estimate	string	3-6, children, adults, n/a, ...	For text data: estimated age of target audience. For speech data: estimated age of speakers.
license	string	cc-by-nc, public domain, ...	The license under which the document is released.
misc	string	{"info": "...", ...}	Optionally included supplementary information as a valid JSON string.
num-tokens	integer	15364	The number of white-separated (or tokenizer-based) tokens present in the document text.
doc-id	string	7a2b3a1d9...	Unique document ID computed as a sha256 string of it's content, used for de-duplication.

Table 3: Document-level schema for the BabyBabelLM datasets. For each document field, we define its type, include sample values, and give a description of its use and contents. The values of the category field are grouped based on which high-level content category they are part of (defined in §3.2.1).

	Language	multiblimp	monoblimp	include	bmlama	multinli	belebele	global-mmlu	arc	hellaswag	xnli	xcopa	xstorycloze	xcomps	winogrande	truthfulqa	sib200
	Random	50.0	50.0	25.0	25.0	33.3	33.3	25.0	25.0	25.0	33.3	50.0	50.0	50.0	50.0	25.0	25.0
Tier 1	Standard Arabic	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Bulgarian	83.5	–	37.5	33.7	57.4	50.6	–	32.8	29.7	37.1	–	52.8	–	51.0	30.8	25.0
	Chinese	–	83.0	52.5	53.3	54.6	61.7	45.8	51.5	37.1	35.5	63.8	63.3	64.5	56.0	26.0	25.0
	Dutch	82.8	–	33.4	40.4	43.0	50.1	37.6	33.2	30.0	–	–	54.0	53.2	51.6	24.0	25.0
	English	98.2	81.0	–	66.5	58.1	55.9	50.0	57.3	40.0	53.4	73.0	67.0	–	55.6	22.4	27.9
	French	96.7	84.6	42.2	46.7	61.0	56.7	38.9	41.4	33.4	44.7	–	58.1	53.0	54.2	26.0	25.5
	German	93.7	86.8	36.0	51.9	33.2	54.0	38.7	37.4	32.3	44.0	–	56.3	54.9	53.4	26.9	25.5
	Indonesian	–	–	44.9	46.3	36.0	53.1	38.8	38.8	32.3	–	57.8	55.7	–	55.8	23.3	25.0
	Persian	81.1	–	33.6	27.9	49.9	46.0	32.6	31.4	29.3	–	–	54.8	52.0	52.8	24.3	25.0
Tier 2	Ukrainian	85.1	–	47.5	33.4	32.7	45.0	34.7	34.4	30.3	–	–	50.9	52.8	52.5	27.6	25.0
	Afrikaans	–	–	–	34.1	37.7	43.8	–	28.3	28.3	–	–	51.2	–	51.4	21.8	25.0
	Basque	89.0	41.6	30.0	–	–	33.9	–	25.4	–	33.5	49.6	49.6	–	–	–	25.0
	Estonian	61.5	–	30.8	23.8	38.1	35.7	–	26.5	27.4	–	47.4	49.3	–	49.3	22.6	25.0
	Greek	92.5	–	33.1	30.7	35.9	42.7	32.5	29.2	29.2	36.3	–	51.9	50.6	52.4	28.4	25.0
	Hebrew	73.0	64.4	42.7	28.3	45.5	40.3	30.9	30.6	28.6	–	–	51.5	50.5	49.2	31.5	25.5
	Italian	88.4	–	45.8	46.4	38.6	53.6	40.2	39.5	33.2	–	56.8	57.0	–	52.4	29.4	25.5
	Japanese	–	74.6	44.7	34.1	58.0	47.1	38.1	39.1	31.4	–	–	56.2	52.8	51.2	27.2	26.0
	Polish	81.5	–	38.1	31.0	39.7	48.6	36.3	34.7	29.8	–	–	53.8	–	53.5	27.6	25.0
	Portuguese	93.4	–	43.2	41.5	38.6	55.2	36.7	41.2	33.9	–	–	58.7	–	53.1	25.5	25.5
	Serbian	–	–	29.3	31.4	41.5	47.8	32.4	28.5	29.6	–	–	53.5	–	52.3	29.2	25.0
	Spanish	93.2	–	42.2	47.7	61.8	48.7	37.6	42.2	34.5	41.4	59.6	58.0	54.3	53.9	28.7	25.5
	Swedish	–	–	–	42.9	53.6	49.3	36.0	32.1	30.3	–	–	52.2	–	50.3	25.7	25.5
	Welsh	70.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Yue Chinese	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	26.0
Tier 3	Achinese	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	25.0
	Balinese	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	25.0
	Buginese	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	25.0
	Croatian	–	–	33.3	32.1	35.9	46.9	–	30.6	29.2	–	–	51.6	–	52.3	28.2	25.0
	Czech	77.8	–	–	30.1	38.8	48.9	36.0	33.5	29.7	–	–	53.1	–	50.8	24.7	25.0
	Danish	92.0	–	–	40.2	36.8	46.4	–	32.0	30.0	–	–	53.8	–	52.4	27.9	25.5
	Hungarian	84.5	–	32.5	26.6	53.1	40.7	–	29.3	28.4	–	–	52.5	51.5	51.1	27.4	25.0
	Icelandic	63.2	–	–	21.7	35.8	35.4	–	26.0	26.7	–	–	46.5	–	49.0	21.9	25.0
	Javanese	–	–	–	36.8	37.0	36.3	–	28.4	27.9	–	–	49.9	–	50.6	19.1	25.0
	Korean	–	–	40.6	32.3	53.1	47.4	35.1	38.0	30.2	–	–	54.3	53.2	51.1	26.0	30.4
	Makasar	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Minangkabau	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	25.0
	Norwegian	–	–	–	40.7	38.5	46.3	–	31.8	29.7	–	–	53.2	–	50.9	24.0	25.0
	Sepedi	–	–	–	–	–	29.6	–	26.1	–	–	–	–	–	–	–	25.0
	Romanian	85.0	–	–	35.4	56.0	51.7	37.3	33.2	30.5	–	–	53.4	–	53.2	25.5	25.0
	Russian	89.6	85.6	44.4	35.9	49.2	53.4	37.9	41.8	33.5	43.7	–	59.4	55.9	53.7	27.4	25.0
	Sesotho	–	–	–	–	–	29.1	27.6	–	–	33.3	–	–	–	–	–	25.0
	Sundanese	–	–	–	–	–	31.6	–	–	–	–	–	–	–	–	–	25.0
	Turkish	79.2	75.4	38.5	32.7	49.0	41.8	34.1	36.3	29.4	38.5	56.2	52.3	51.3	48.6	26.9	25.0
	isiXhosa	–	–	–	–	–	28.7	26.2	–	–	33.3	–	–	–	–	–	25.0
	isiZulu	–	–	–	–	–	32.0	25.0	26.3	–	33.3	–	–	–	–	–	25.0

Table 4: Performance of Qwen3-0.6B (Yang et al., 2025). All scores denote average 0-shot accuracy. Columns are sorted by the column order of Table 1.

	Language	multiblimp	monoblimp	bmlama	xcopa	arc	xcomps	hellaswag	xnli	multinli	winogrande	sib200	belebele	global-mmlu	include	xstorycloze	truthfulqa
	Random	50.0	50.0	10.0	50.0	25.0	50.0	25.0	33.3	33.3	50.0	14.3	25.0	25.0	25.0	50.0	25.0
TIER 1 (100M)	Bulgarian	90.8	–	24.7	–	25.3	–	26.8	34.5	32.0	52.0	19.6	28.3	–	28.0	49.5	25.5
	Chinese	–	70.2	34.5	51.2	36.9	55.1	26.8	33.3	35.8	49.2	11.3	22.3	23.0	23.9	48.7	22.1
	Dutch	90.5	–	29.4	–	30.3	52.4	26.5	–	34.6	50.0	9.3	27.0	24.5	32.4	49.1	22.4
	English	82.0	65.9	31.8	58.0	28.8	–	26.5	36.3	31.8	51.4	24.0	23.0	23.2	–	49.5	22.4
	French	94.1	69.7	22.7	–	27.5	50.6	26.4	36.1	35.7	51.3	8.8	27.3	23.9	22.7	47.6	20.4
	German	88.6	77.1	26.3	–	27.1	52.6	25.9	34.9	32.3	51.8	11.8	24.7	24.6	16.3	48.8	23.8
	Indonesian	–	–	25.6	53.6	30.4	–	27.3	–	31.7	53.1	10.8	25.4	26.1	22.9	50.5	24.3
	Persian	71.3	–	29.6	–	29.7	53.6	26.4	–	35.3	50.7	9.8	26.0	23.8	23.0	52.2	26.9
	Ukrainian	88.6	–	24.8	–	28.4	50.6	26.4	–	31.6	50.3	8.3	22.8	23.4	30.6	47.6	24.7
TIER 2 (10M)	Afrikaans	–	–	28.5	–	28.3	–	26.1	–	32.0	51.3	10.8	22.1	–	–	49.5	22.6
	Arabic	75.9	–	17.2	–	25.7	52.9	25.8	33.0	32.4	47.4	19.1	22.9	23.1	22.1	46.8	22.3
	Basque	94.5	65.3	–	49.8	28.5	–	–	33.6	–	–	10.3	26.0	–	25.4	50.6	28.0
	Estonian	81.5	–	21.0	53.0	24.8	–	25.5	–	35.9	50.7	12.2	22.0	–	21.4	45.8	22.6
	Greek	89.2	–	23.2	–	26.9	50.3	26.4	35.6	31.7	49.5	19.6	22.8	23.1	20.9	49.4	25.7
	Hebrew	70.2	59.5	23.0	–	29.8	51.6	26.1	–	32.5	50.2	12.2	22.9	23.1	24.8	49.6	23.1
	Italian	77.5	–	26.3	52.0	26.4	–	26.5	–	31.7	50.1	11.3	24.6	23.3	20.7	50.2	25.2
	Japanese	–	61.9	19.9	–	27.8	50.8	25.1	–	31.7	47.5	19.6	23.4	23.0	26.1	47.8	22.1
	Polish	75.9	–	19.4	–	26.1	–	25.5	–	32.0	49.0	12.2	21.9	26.6	18.2	49.5	21.8
	Portuguese	80.7	–	21.7	–	25.4	–	26.3	–	31.7	48.8	12.2	23.9	23.5	22.2	48.8	23.6
	Serbian	–	–	24.4	–	25.5	–	25.6	–	32.4	49.5	8.3	22.9	23.1	20.2	48.5	23.5
	Spanish	83.0	–	24.1	–	28.0	51.3	26.4	35.5	33.7	49.1	18.6	24.9	24.4	23.7	47.8	23.0
	Swedish	–	–	23.3	–	26.2	–	25.9	–	31.7	49.3	12.2	28.3	24.6	–	48.1	24.5
	Welsh	91.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Yue Chinese	–	–	–	–	–	–	–	–	–	–	19.6	–	–	–	–	–
TIER 3 (1M)	Achinese	–	–	–	–	–	–	–	–	–	–	18.1	–	–	–	–	–
	Balinese	–	–	–	–	–	–	–	–	–	–	23.5	–	–	–	–	–
	Buginese	–	–	–	–	–	–	–	–	–	–	19.6	–	–	–	–	–
	Croatian	–	–	17.8	–	26.1	–	25.8	–	31.7	47.6	12.2	22.4	–	25.4	45.6	21.1
	Czech	59.0	–	13.1	–	24.8	–	25.8	–	35.9	50.4	19.6	22.9	23.1	–	48.6	20.2
	Danish	82.0	–	17.2	–	24.6	–	25.7	–	35.9	49.0	12.2	23.6	–	–	48.9	21.8
	Hungarian	68.9	–	10.2	–	24.3	49.9	25.6	–	31.7	48.0	12.2	22.9	–	30.4	48.5	18.9
	Icelandic	71.6	–	13.4	–	24.3	–	25.4	–	31.7	50.0	12.2	22.9	–	–	45.7	22.6
	Javanese	–	–	17.4	–	25.2	–	25.8	–	31.7	50.5	19.6	28.2	–	–	50.6	24.1
	Korean	–	–	19.1	–	25.5	51.3	25.0	–	32.5	48.9	10.8	21.6	26.5	21.8	47.5	23.3
	Makasar	–	–	–	–	–	–	–	–	–	–	12.4	–	–	–	–	–
	Minangkabau	–	–	–	–	–	–	–	–	–	–	19.6	–	–	–	–	–
	Norwegian	–	–	21.4	–	26.4	–	25.9	–	35.9	47.0	12.2	22.9	–	–	47.1	22.3
	Sepedi	–	–	–	–	25.9	–	–	–	–	–	10.8	24.1	–	–	–	–
	Romanian	74.1	–	15.4	–	25.3	–	25.5	–	35.7	48.6	14.7	24.0	24.4	–	46.2	21.8
	Russian	58.5	52.0	17.2	–	25.2	49.1	25.9	33.3	35.9	51.3	12.2	28.9	24.9	20.5	48.8	21.4
	Sesotho	–	–	–	–	–	–	–	33.3	–	–	9.3	22.9	20.0	–	–	–
	Sundanese	–	–	–	–	–	–	–	–	–	–	19.6	27.2	–	–	–	–
	Turkish	64.9	59.8	17.9	53.8	25.4	51.6	25.9	33.3	35.9	49.9	12.2	21.9	27.1	23.0	50.1	24.7
	isiXhosa	–	–	–	–	–	–	–	33.3	–	–	10.8	22.9	20.0	–	–	–
	isiZulu	–	–	–	–	29.7	–	–	33.3	–	–	10.8	23.0	22.8	–	–	–

Table 5: Zero-shot performance across all tasks of the monolingual GPT-2 models trained on BabyBabelLM. Columns are sorted by difference of the average task performance against random chance.

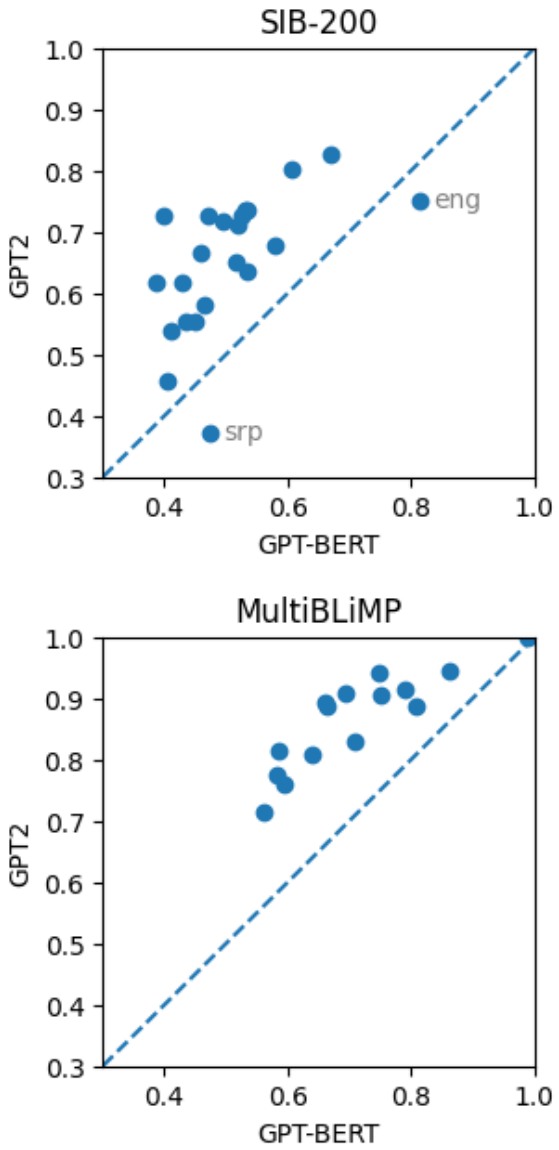

Figure 4: GPT-2 and GPT-BERT accuracy scores on SIB-200 and MultiBLiMP.