TafBERTa: Learning Grammatical Rules from Small-Scale Language Acquisition Data in Hebrew

Anita Gelboim[1], **Elior Sulem**[1,2]
[1]Faculty of Computer and Information Science, Institute for Applied AI Research
[2]The School of Brain Sciences and Cognition
Ben-Gurion University of the Negev
anitana@post.bgu.ac.il, eliorsu@bgu.ac.il

Abstract

We present TafBERTa, a compact RoBERTa (Liu et al., 2019) based language model tailored for Hebrew child-directed speech (CDS). This work builds upon the BabyBERTa (Huebner et al., 2021) framework to address data scarcity and morphological complexity in Hebrew. Focusing on determiner-noun grammatical agreement phenomena, we show that TafBERTa achieves competitive performance compared to large-scale Hebrew language models while requiring significantly less data and computational resources. As part of this work, we also introduce a new corpus of Hebrew CDS, HTBerman, aligned with morphological metadata and our new grammatical evaluation benchmark for Hebrew, HeCLiMP, based on minimal pairs. Our results demonstrate the effectiveness of TafBERTa in grammaticality judgments and its potential for efficient NLP in low-resource settings.

1 Introduction

In the last few years, Language Models (LMs) have expanded in both parameter count and training data size (Kaplan et al., 2020). Besides the numerous contributions to NLP tasks (Min et al., 2023; Zhao et al., 2023) and their application in many domains (Chiarello et al., 2024), this trend brings various challenges, including computational inefficiency, increased environmental costs and difficulties in adapting models to low-resource languages.

Recently, works such as BabyBERTa (Huebner et al., 2021) and the BabyLM Challenge (Warstadt et al., 2023) addressed these aspects by developing English compact models trained on child-directed language, demonstrating strong grammatical abilities with minimal data. However, no such efforts have not been done in Hebrew, a low resource language where data scarcity is a main challenge, leaving a significant gap in efficient, accessible language modeling for Hebrew NLP.

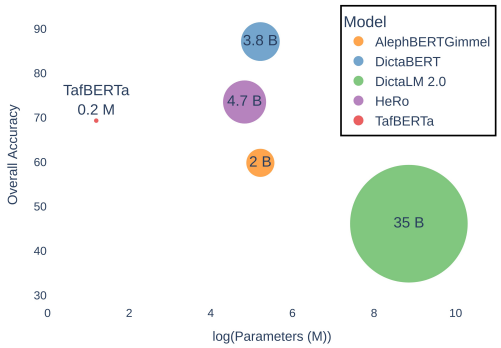

Figure 1: Overall accuracy of Hebrew language models on the HeCLiMP benchmark (see Section 4). Bubble size represents the number of words seen during training, while the x-axis indicates the logarithm of model parameters (M).

In this paper, we introduce TafBERTa, a compact RoBERTa (Liu et al., 2019) based model optimized for Hebrew. To assess the effectiveness and efficiency of TafBERTa, we pose several key research questions. First, we investigate how TafBERTa's smaller size—defined by both its reduced number of parameters and the smaller dataset used for training—impacts its performance relative to HeRo (Shalumov and Haskey, 2023, a Hebrew version of RoBERTa). This comparison assesses whether a more compact architecture can achieve competitive results, despite having fewer computational resources and less training data (Q1). Beyond this direct comparison, we explore whether a search over the parameter space was necessary for optimizing TafBERTa's performance, particularly in training a RoBERTa architecture on the HTBerman child-directed speech corpus we introduce (Q2). Additionally, we evaluate the capabilities of TafBERTa against other Hebrew models using other architectures or tokenization methods, to establish its relative strengths and weaknesses within

Proceedings of the First BabyLM Workshop, pages 76–90

the Hebrew NLP landscape (Q3). Finally, we assess the adaptability of TafBERTa's architecture by testing its ability to learn from alternative data sources, specifically evaluating its performance when trained on Wikipedia-derived Hebrew text rather than child-directed speech (Q4). These questions guide our evaluation, providing insights into both the efficiency of small-scale models and the nuances of Hebrew NLP.

Our contributions: (1) introducing TafBERTa, an efficient Hebrew model, (2) introducing HTBerman dataset for Hebrew Child-Directed Speech (CDS), (3) presenting HeCLiMP, a benchmark for Hebrew grammatical evaluation tailored to CDS, and (4) conducting a comparative study against HeRo and other models. Results show TafBERTa achieves competitive performance despite its reduced size, highlighting the potential of small, well-tuned models for low-resource NLP.[1]

2 Related Work

2.1 Baby Language Models

In response to the parameter and data expansion in Large Language Models (LLMs), research has increasingly turned toward smaller, more efficient models that retain strong linguistic capabilities. The BabyLM challenge (Warstadt et al., 2023) exemplifies this shift, encouraging the development of compact models that learn from limited yet high-quality data, mimicking human language acquisition. A key resource is the CHILDES database (MacWhinney, 2000), which includes well-established corpora of casual speech to children that has shaped studies in cognitive linguistics and NLP (Huebner and Willits, 2021; Mueller and Linzen, 2023). Building on this foundation, BabyBERTa (Huebner et al., 2021) was introduced as a scaled-down RoBERTa variant trained on child-directed language, demonstrating that even with fewer parameters and less training data, models can develop strong grammatical abilities. Evaluation of such models relies on syntactic and grammatical benchmarks like BLiMP (Warstadt et al., 2020) and Zorro (Huebner et al., 2021), which test linguistic phenomena.

We address here these questions from the perspective of the Hebrew language, tackling challenges in low-resource language adaptation.

[1]We release the code and datasets at https://github.com/NLU-BGU/tafberta/ to facilitate reproducibility and future research.

2.2 Baby Language Models in Other Languages

While much of the research on baby language models has focused on English, recent work has expanded these efforts to additional languages.
For instance, in Italian, Capone et al. (2024) introduced a benchmark designed for the standardized evaluation of Italian BabyLMs. To assess its effectiveness, researchers applied the benchmark to Minerva (Orlando et al., 2024), an LLM pretrained from scratch on Italian. The results revealed that Minerva struggled with certain linguistic aspects, achieving an age-equivalent score of just four years. This under-performance highlights the necessity of refining model training approaches to improve language acquisition efficiency. In German, Bunzeck et al. (2025) studied the effect of utterance-level construction distributions in German child-directed and child-available speech on the model performance at the word-level, syntactic and semantic levels. The grammatical abilities of Baby Language Models beyond English have also been investigated in Salhan et al. (2024), covering Chinese, French, German, and Japanese, and focusing on the effect of curriculum learning. Focusing on phonology, Goriely and Buttery (2025) trained small monolingual language models on child-directed and child-produced speech, covering 11 languages.

Several recent studies have explored second language acquisition (L2) with language models, drawing parallels to human language learning processes. In Italian, BAMBINO-LM (Shen et al., 2024), a bilingual pre-training approach for BabyLM, enhances Italian proficiency while maintaining English skills, using alternation and PPO (proximal policy optimization)-based perplexity rewards. Yadavalli et al. (2023) and Oba et al. (2023) examined L2 acquisition in neural models, by pretraining LMs in a certain language, further training them in English as an L2, and evaluating and analyzing their linguistic generalization in L2. They found that L1 pretraining accelerates L2 learning, with varying linguistic transfer effects.

We focus here on Hebrew, a low-resource language for which Baby Language Models have not been explored, and address it in a monolingual setting.

2.3 Hebrew Language Models

Hebrew language models continue to lag behind their English counterparts, facing challenges in data

availability and computational efficiency (Tsarfaty et al., 2019).

Compared to English, Hebrew has more limited corpora for training large-scale models, making it difficult to achieve the same level of performance. Despite this limitation, several Hebrew language models have been developed to bridge the gap. AlephBERT (Seker et al., 2022), AlephBERTGimmel (Gueta et al., 2023) and HeRo (Shalumov and Haskey, 2023) were among the first transformer-based models for Hebrew providing contextual embeddings suited to the language's structure. DictaBERT (Shmidman et al., 2023) and its successor, DictaLM 2.0 (Shmidman et al., 2024), further refined Hebrew language modeling, improving general-purpose NLP tasks. While these advancements mark progress, Hebrew NLP still requires larger, higher-quality datasets and more efficient training strategies to reach the capabilities of English LLMs.

Another difficulty in Hebrew NLP is the linguistic challenge (Tsarfaty et al., 2019). Hebrew is a morphologically-rich language (MRL), and in MRLs, every input token could contain some lexical and functional units, known as morphemes, each playing a distinct role in shaping the syntactic or semantic representation. One challenge arises from the necessity to segment Hebrew tokens into their constituent morphemes before processing Hebrew texts. The segmentation process has experienced significant advancement with the utilization of tools like YAP (Yet Another (Natural Language) Parser, More et al., 2019) or the DictaBERT model (Shmidman et al., 2023), which has been fine-tuned specifically for the segmentation task.

In Hebrew NLP, only after performing the segmentation phase, we should chose the tokenizer. The most popular tokenization algorithms are Byte-Pair Encoding (BPE) (Sennrich et al., 2016) and Google's WordPiece (Song et al., 2021), which are used by RoBERTa and BERT respectively. Another method based on morphemes is used by HeBERT and AlephBERTGimmel. See Gazit et al. (2025) and Gorman and Pinter (2025) for further perspectives on Hebrew tokenization.

Our work takes these challenges into account by focusing on data efficiency and morphological complexity, designing a model that learns from limited yet high-quality Hebrew data while addressing the constraints of low-resource language modeling.

2.4 Probing Grammatical Rule Learning

Recent LLMs have demonstrated remarkable success in addressing a wide array of downstream tasks. However, there is still a need to determine the extent to which these LLMs comprehend the syntax of natural languages. To tackle this question, several studies have examined the syntactic understanding of language models using tailored datasets specifically designed for targeted syntactic evaluations. One way to examine it is using a probing task i.e., a classification problem that focuses on simple linguistic properties of sentences (Conneau et al., 2018). The objective of this task is to assess the quality of a model, focusing on its language proficiency, particularly in syntax and grammar. Some explored this question by evaluating language models' (LMs) preferences between *minimal pairs* (MP) of sentences differing in grammatical acceptability, as in the next example:

1. Imagination is more important than knowledge. (grammatical)

2. Imagination are more important than knowledge. (ungrammatical)

A MP is classified correctly if a LM assigns a higher probability to the grammatical sentence than to the ungrammatical one.
The Benchmark of Linguistic Minimal Pairs (BLiMP, Warstadt et al., 2020) is a benchmark designed with linguistic principles in mind. It evaluates the ability of language models to discern acceptability differences across various English phenomena. However, most of the studies have focused on English and other European languages. Only few studies extended this investigation to non-European languages, such as CLIMP (Xiang et al., 2021) and JBLiMP (Someya and Oseki, 2023), for Chinese and Japanese languages respectively. The authors of CLIMP built the corpus of Chinese MPs in the by generating data from grammar templates for every paradigm they incorporate, building an annotated vocabulary, and generating sentences by sampling words from the vocabulary, which is a translation of BLIMP English Vocabulary. The authors of JBLIMP created the corpus of Japanese MPs based on acceptability judgments extracted from journal articles in theoretical linguistics. These minimal pairs are grouped into 11 categories, each covering a different linguistic phenomenon. In some other languages (specifically Italian, English, Hebrew and Russian), Gulordava

et al. (2018) strengthen the evaluation paradigm of MPs in terms of subject-verb agreement. Their assessment involves nonsensical sentences, challenging language models by eliminating reliance on semantic or lexical cues ("The colorless green ideas I ate with the chair sleep furiously"). The evaluation test sets are extended to other phenomena, resulting in the CLAMS benchmark (Mueller et al., 2020).

Differently from the Hebrew section of CLAMS, we build here a grammatical benchmark (HeCLiMP) tailored to CDS, abstracting away from lexical complexity, yet addressing two main Hebrew grammatical phenomena that exemplify the rich morphology in Hebrew. HeCLiMP also differs from CLAMS by being constructed directly in Hebrew, abstracting away from the effects of translation from the English language.

Grammatical benchmarks in English that are tailored to CDS include Zorro (Huebner et al., 2021) and BabySLM (Lavechin et al., 2023).

3 Training Data: HTBerman dataset

Our main corpora of interest are the original version of CHILDES Hebrew Berman Longitudinal Corpus (Armon-Lotem, 1996, Berman corpus)[2], written in latin-based phonemic Hebrew talk transcription and a version of it written in standard Hebrew script (Albert et al., 2012).[3]

The Berman corpus comprises longitudinal naturalistic data gathered weekly from four Hebrew-speaking children. In order to fairly compare with other Hebrew language models, we use the version of the corpus written in standard Hebrew script. Since the latter does not contain the metadata present in the original version, we merge the two versions, creating a comprehensive dataset that incorporates Hebrew text along with all the annotations, at the utterance and word levels.

As part of the corpora merge, we performed data cleaning, which included morphological segmentation (More et al., 2019) and punctuation correction (See Section A for the details). The resulted corpus **HTBerman** (Hebrew Transcription Berman) contains 53K sentences, 233k words and ∼8K unique words of Hebrew transcribed CDS.

[2]This corpus is part of CHILDES project (MacWhinney, 2000).

[3]We use as initial data the outputs of the automatic converter built by Albert et al. (2012).

3.1 Corpora

3.1.1 CHILDES Hebrew Berman Longitudinal Corpus

Our main corpus of interest is CHILDES Hebrew Berman Longitudinal Corpus. This corpus is transcribed with a Latin-based phonemic of Hebrew talks. The transcripts were all transcribed in the CHAT format (CHILDES) with adaptations to Hebrew. The dataset comprises longitudinal naturalistic data gathered weekly from four Hebrew-speaking children. These children are all native Hebrew speakers raised in households where Hebrew is the primary language and the environment is characterized by high levels of education. Each child was audio-recorded in various settings at their home, including mealtime, bath time, solitary play and interactions with siblings, parents and grandparents. This corpus includes the following morphological annotations:

- **Participants:** This component refers to the individuals involved in the conversation. In CHILDES, the convention is to designate the child being studied as CHI and the child's mother as MOT. Each utterance in the conversation begins with an indication of the participant speaking, denoted by an asterisk (*) followed by the participant code.
- **Transcriptions:** Transcriptions capture the spoken language in written form.
- **Dependent tiers:** These are additional layers of linguistic information associated with each transcription line. They are preceded by a percentage symbol (%) and are linked to the transcription line immediately above. Dependent tiers can include morphological information (%mor), grammatical relations (%gra), intonation (%int) and others. While some tiers are common in CHILDES datasets, none are obligatory.
- **The %mor tier:** This tier provides morphological information about each word in the transcription. It aligns one-to-one with the segmented words and disregards any annotations present in the transcription line. Each item in the %mor tier consists of a part-of-speech tag followed by inflectional or derivational information, separated by a pipe (|). For example, "qn|more" indicates a nominal quantifier aligned with the word "more".

- **The %gra tier:** The grammatical relations tier represents relationships between words in terms of heads and dependents in dependency grammar. Each item in the %gra tier corresponds one-to-one with the segmented words in the transcription, as well as with items in the %mor tier. It specifies the syntactic relationship between words, such as subject-verb or quantifier-noun.

- **Other tiers:** In addition to %mor and %gra, there may be other dependent tiers providing further linguistic or contextual information. For example, the %int tier captures intonation patterns, while others may contain information about the recording session or the context of the conversation.

We accessed the data using PyLangAcq (Lee et al., 2016).

3.1.2 Standard Hebrew Berman Longitudinal CHILDES Corpus

The Standard Hebrew Berman Longitudinal CHILDES corpus has the same talks as in 3.1.1, but written in standard Hebrew. This corpus has only raw data of Hebrew text, while the original one, transcribed in latin-based phonemic, has also morphological annotations as metadata.
Our objective is to merge these datasets, creating a comprehensive dataset that incorporates Hebrew text along with all the annotations from 3.1.1, both at the utterance level and the word level. The annotations are needed for the creation of the HeCLIMP evaluation benchmark (See Section 4).

3.2 Corpora Merge and Data Preprocessing

The corpora merge involves file-level, utterance-level, and token-level matching. As part of the corpora merge, we performed data cleaning, which included morphological segmentation (More et al., 2019) and punctuation correction (See Appendix A for more details). The resulted corpus **HTBerman** (Hebrew Transcription Berman) contains 53K sentences, 233k words and ∼8K unique words of Hebrew transcribed CDS.

4 HeCLiMP Evaluation Benchmark

We compile HeCLiMP (Hebrew Child-Directed Linguistic Minimal Pairs), a Hebrew CDS grammar test suite, to evaluate how well language models grasp grammaticality in an environment that closely reflects the linguistic input children receive.

Based on minimal pairs (Conneau et al., 2018), HeCLiMP is composed of sentence pairs that differ by just one key element — one sentence is grammatically correct and the other is minimally incorrect. We focus on two grammatical phenomena, adapting Determiner-Noun (DN) agreement from BLiMP and Zorro to Hebrew. By doing so, we address a phenomenon that exists in English (number agreement) and one that does not hold in English (gender agreement):
(1) **DN Number Agreement**: e.g., *'ha-kova ha-ze' ('this hat'-singular) vs. 'ha-kovaim ha-ele' ('these hats'-plural).*
(2) **DN Gender Agreement**: Unlike English, Hebrew requires determiners to match the gender of the noun , e.g., *'ha-kova ha-ze' ('this hat'-masc.) vs. 'ha-simla ha-zo' ('this dress'-fem.).*

Following the procedure used for Zorro in the case of English, we generated minimal pairs using template filled with words from HTBerman (Section 3). Each paradigm consists of 5,596 minimal pairs in the test set and 1,398 minimal pairs in the development set.

Most existing grammar evaluation benchmarks in NLP focus on adult-directed language, posing challenges for assessing the grammatical competence of models trained on CDS. To address this gap in the case of Hebrew, we developed HeCLiMP, a benchmark specifically designed to evaluate Hebrew grammatical learning in models trained on CDS. Our approach follows the methodology of BLiMP and Zorro, but with simplified templates that prioritize morphological features relevant to Hebrew language acquisition.

To construct test sentences, we first designed a set of sentence templates for each grammatical paradigm. These templates were then populated with words sampled from HTBerman, ensuring that all inserted content words conformed to the necessary morphological constraints. Word lists were generated by filtering nouns from HTBerman along with their gender and number annotations.

A primary focus of HeCLiMP is determiner-noun agreement in Hebrew, specifically gender and number agreement. We used simple templates such as *"Look at this ..."* or *"Look at that ..."*, where the determiner adapted according to the gender and number of the noun.

	HeRo	TafBERTa
Parameters	125M	3.3M
Data size	47.5GB	1.8MB
Words in data	4.7B	233k
Batch size	8k	128
Max sequence	512	128
Epochs	25	5
Hardware	1xGTX1080	1xRTX6000
Training time	35 days	105 seconds
Model Configurations		
Vocabulary size	50K	7317
Hidden size	768	64
Layers number	12	10
Attention heads	12	4
Intermediate size	3072	2048
Max. sequence	512	128
Accuracy *	73.5	69.4

Table 1: A Comparison between HeRo, pre-trained on 4.7B words of web text, and TafBERTa, pre-trained from scratch on 233k words of child-directed input. *Accuracy results on the evaluation task.

5 TafBERTa

Model We introduce a scaled-down masked language model based on RoBERTa, with 3.3M parameters, 7317 vocabulary items trained on 233K words. We will refer to this model as TafBERTa[4]. All hyper-parameters were identified by tuning TafBERTa on a masked word prediction task using a held-out portion of our corpus of transcribed CDS as input. A detailed comparison between hyper-parameters of TafBERTa and other Hebrew LMs we compared to is in Tables 1 and 2. Briefly, TafBERTa uses only 10 layers, 4 attention heads, 64 hidden units and an intermediate size of 2048.

Vocabulary TafBERTa uses a Byte-Pair Encoding (Sennrich et al., 2016) sub-word vocabulary, like HeRo and RoBERTa. Instead of HeRo's 50K word vocabulary, we built a 7317-word vocabulary from HTBerman.

Hyper-Parameters Search We optimized hyper-parameters on the development set (Table 5), focusing on those with significant improvements in BabyBERTa. The development set consisted of the two DN agreement paradigms from HeCLiMP.

6 Experiments

Results reflect the average performance over six runs with different seeds including RoBERTa on HTBerman (§6.2), BabyBERTa on HTBerman (§6.2), and the Wikipedia-trained model (§7).

The models used for the comparison are HeRo [5], AlephBERT [6], AlephBERTGimmel [7], DictaBERT [8], and DictaLM2.0 [9]. Differently from the other models, which are encoder-based language models, DictaLM2.0 is a large decoder-based language model. A comparison between the models is presented in Table 2.

6.1 Evaluation Method

Inspired by the BabyBERTa paper, we use *holistic scoring* (Zaczynska et al., 2020). For each minimal pair, we calculate the model's preference for the grammatical sentence over the ungrammatical one. This score is obtained by summing the cross-entropy errors across all positions in the sentence. Accuracy is the ratio of correct choices to total pairs.

6.2 Results

The results are presented in Table 3. These questions (Q1, etc.) are as described in the introduction.

[4]Taf means toddler in Hebrew

[5]https://huggingface.co/HeNLP/HeRo

[6]https://huggingface.co/onlplab/alephbert-base

[7]https://huggingface.co/imvladikon/alephbertgimmel-base-512

[8]https://huggingface.co/dicta-il/dictabert

[9]https://huggingface.co/dicta-il/dictalm2.0

Model	AlephBERT	AlephBERTGimmel	DictaBERT	DictaLM 2.0	HeRo	TafBERTa
Parameters	126M	184M	184M	7B	125M	**3.3M**
Words in data	1.9B	2B	3.8B	35B*	4.7B	**233K**

Table 2: Comparison of model sizes and training data.

Model	Overall	Number	Gender	Tokenizer	Model
AlephBERT	58.6	58.7	58.5	WP	Encoder
AlephBERT-Gimmel	59.8	54.3	65.3		
DictaBERT	87.1	90.1	84.2		
DictaLM2.0	46.1	31.4	60.8	BPE	Decoder
HeRo	73.5	69.1	77.9	BPE	Encoder
RoBERTa (HTBerman)	65.6	83.7	47.5		
TafBERTa	69.4	80.5	58.2		

Table 3: Accuracy on each phenomenon in HeCLiMP. We used the Holistic-scoring method. “Overall” refers to the overall accuracy across all phenomena. “Number” and “Gender” refer to determiner-noun agreement in number and gender, respectively. WP refers to the WordPiece tokenizer.

Model	Overall	Number	Gender
Wikipedia	43.3	30.9	55.7
TafBERTa	**69.4**	**80.5**	**58.2**

Table 4: Performance of the Wikipedia-Trained Model and TafBERTa on the HeCLIMP subset. "Overall" refers to the overall accuracy across all phenomena. “Number” and “Gender” refer to determiner-noun agreement in number and gender, respectively. The highest score in each column is highlighted in bold.

Comparison with HeRo (Q1) Since HeRo and TafBERTa share the same architecture and tokenizer, the comparison between the two allows for a direct assessment of the impact of training data and optimization choices. TafBERTa achieved an overall accuracy of 69.4 on the test set while HeRo reaches 73.5. Breaking this down by task, we observe an interesting tradeoff: while TafBERTa excels in DN agreement for number (80.5 vs. 69.1), HeRo demonstrates superior performance in DN agreement for gender (77.9 vs. 58.2).

Comparison to RoBERTa trained on HTBerman (Q2) We trained the RoBERTa architecture on HTBerman using the same number of epochs as TafBERTa. It achieved 65.6 overall accuracy, with strong performance on number agreement (83.7) but poor results on gender agreement (47.5). This highlights the importance of tailored pre-training objectives and hyperparameter optimization, as seen in TafBERTa, to achieve balanced performance across linguistic tasks. A further analysis of RoBERTa is presented in Appendix D.

Comparison to BabyBERTa TafBERTa and BabyBERTa share the same underlying architecture, but differ in their hyper-parameters. To directly compare the two, we trained BabyBERTa's architecture using its original hyper-parameters on HTBerman. BabyBERTa achieved lower performance than TafBERTa on the two tasks, suggesting that careful adaptation of hyper-parameters is crucial when applying a shared architecture to different languages.

Additional comparisons (Q3) We observe that DictaLM 2.0, a Large Language Model being the current state-of-the-art (SOTA) for Hebrew in general tasks, performed the worst on the number agreement task, achieving only 31.4 accuracy, sig-

nificantly below other models.

In the group of RoBERTa-based models using WordPiece tokenizers, DictaBERT achieved the highest overall accuracy in this group (87.1), with especially strong results in number agreement (90.1). In contrast, AlephBERT and AlephBERT-Gimmel lagged behind, with overall accuracies of 58.6 and 59.8, respectively, reflecting less robust handling of grammatical tasks.

7 Alternative Training Data

We assess the adaptability of TafBERTa's architecture by testing its ability to learn from alternative data sources, specifically evaluating its performance when trained on Wikipedia-derived Hebrew text rather than CDS (**Q4**). We utilized the SVLM Hebrew Wikipedia Corpus[10], preprocessed in the same manner as HTBerman. The dataset size was adjusted to match the word count of HTBerman, ensuring equivalent scales for training.

Using this dataset, we trained a new language model that retained the architecture of TafBERTa but replaced the training data with the processed Wikipedia corpus. Subsequently, we evaluated this new model on a subset of HeCLIMP, focusing on minimal pairs containing words seen by the model during training. For comparison, we also assessed TafBERTa on the same test set. That is to say, the two models we compare have seen during training the words used in the benchmark, and only differ by the type of the training data used (HTBerman vs. Wikipedia).

The results (Table 4) indicate that while the Wikipedia corpus serves as a rich and diverse resource, its effectiveness in training for grammatical agreement tasks is limited compared to the original dataset used for TafBERTa.

8 Conclusion

We present in this paper TafBERTa, a first language model tailored to Hebrew Child-Directed Speech. Focusing on Determiner-Noun agreement phenomena, we show that TafBERTa shows competitive performance with larger Hebrew language models. By doing so, we extend acquisition-inspired, small-scale language model research to a low-resource language, where such efforts are particularly needed. Our results emphasize the need for language-specific and data-specific tuning to fully leverage the capabilities of such models. Future work includes the extension of HeCLiMP to additional grammatical phenomena, the use of training data originated from later stages of language acquisition (i.e., language directed to older children), and the exploration of alternative language model architectures.

[10] https://github.com/NLPH/SVLM-Hebrew-Wikipedia-Corpus

Limitations

While TafBERTa demonstrates progress in modeling Hebrew child-directed speech, several limitations highlight areas for future work and improvement.

Evaluation Improvements Our evaluation framework, HeCLiMP, successfully benchmarks grammatical proficiency but remains limited in scope. Currently, it focuses on determiner-noun agreement in gender and number. Future work should expand HeCLiMP to include a set of grammatical structures, such as verb-subject agreement and determiner-noun agreement with an adjective in between.

Multilingual Model Development While TafBERTa is optimized for Hebrew, its application is restricted to a monolingual context. Extending the model to a multilingual framework by training on related Semitic languages (e.g., Arabic) could enhance its ability to generalize across linguistic variations.

Training on Older Children's Data Currently, TafBERTa is trained on speech data directed at younger children, which captures early-stage language acquisition patterns. However, language complexity increases with age. Training on speech data directed at older children would enable the model to learn more advanced syntactic and morphological structures, better simulating additional phases of language development.

Exploring Alternative Architectures The BERT architecture has dominated Hebrew NLP research and TafBERTa follows this trend. However, exploring other architectures may yield performance improvements. Additionally, architectures optimized for low-resource settings, such as efficient transformers (e.g., DistilBERT (Sanh et al., 2020)), could offer a better trade-off between computational efficiency and linguistic expressiveness.

Ethics Statement

The language acquisition data we are using in this work were taken from the TalkBank system[11], with includes CHILDES, and where all contributions have received an IRB approval. Our own work on the data has been approved by the Ben-Gurion University of the Negev Ethics committee.

Acknowledgements

We would like to thank Shuly Wintner and Bracha Nir for sharing with us the CHILDES data converted into Hebrew script, which forms the basis of the HTBerman corpus presented in this work, and the anonymous reviewers for their helpful comments. We also acknowledge the NICHD HD082736 grant support for CHILDES. Our work was supported in part by grants from the Israeli Ministry of Innovation, Science & Technology (#000519) and from the Data Science Research Center at Ben-Gurion University of the Negev.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Aviad Albert, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2012. A morphologically annotated hebrew childes corpus. In *Proc. of the Workshop on Comp. Models of Language Acquisition and Loss*, pages 20–22.

Sharon Armon-Lotem. 1996. *The minimalist child: Parameters and functional heads in the acquisition of Hebrew*. Tel Aviv University.

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.

Luca Capone, Alice Suozzi, Gianluca Lebani, and Alessandro Lenci. 2024. Babies: A benchmark for the linguistic evaluation of italian baby language models. In *Italian Conf. on Comp. Ling.*

Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, and Gualtiero Fantoni. 2024. Future applications of generative large language models: A data-driven case study on chatgpt. *Technovation*, 133:103002.

[11] https://childes.talkbank.org/

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Bar Gazit, Shaltiel Shmidman, Avi Shmidman, and Yuval Pinter. 2025. Splintering nonconcatenative languages for better tokenization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22405–22417, Vienna, Austria. Association for Computational Linguistics.

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.

Kyle Gorman and Yuval Pinter. 2025. Don't touch my diacritics. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 285–291, Albuquerque, New Mexico. Association for Computational Linguistics.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2023. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. ArXiv 2211.15199 [cs.CL].

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proc. of NAACL'18*, pages 1195–1205.

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.

Philip A. Huebner and Jon A. Willits. 2021. Chapter eight - using lexical context to discover the noun category: Younger children have it easier. In Kara D. Federmeier and Lili Sahakyan, editors, *The Context of Cognition: Emerging Perspectives*, volume 75, pages 279–331. Academic Press.

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.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*. ISCA.

Jackson L. Lee, Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. 2016. Working with chat transcripts in python. Technical report, Department of Computer Science, University of Chicago.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk, Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Amir More, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from Modern Hebrew. *Transactions of the Association for Computational Linguistics*, 7:33–48.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada. Association for Computational Linguistics.

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.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. In *Italian Conf. on Comp. Ling.*

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56, Dublin, Ireland. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Vitaly Shalumov and Harel Haskey. 2023. Hero: Roberta and longformer hebrew language models. *arXiv preprint arXiv:2304.11077*.

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.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities. *Preprint*, arXiv:2407.07080.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. ArXiv 2308.16687 [cs.CL].

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.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proc. of EMNLP'21*, pages 2089–2103.

Reut Tsarfaty, Shoval Sadde, Stav Klein, and Amit Seker. 2019. What's wrong with Hebrew NLP? and how to make it right. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pages 259–264, Hong Kong, China. Association for Computational Linguistics.

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. *TACL*, 8:377–392.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proc. of EACL*, pages 2784–2790.

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.

Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. Evaluating German transformer language models with syntactic agreement tests. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020*, volume abs/2007.03765, Zurich, Switzerland. CEUR Workshop Proceedings.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. ArXiv 2303.18223 [cs.CL].

A Data Preprocessing for the HTBerman Construction

Our ultimate aim is to train TafBERTa using Hebrew Child-Directed Speech data. To accomplish this, we must filter the CDS utterances in Standard Hebrew corpus 3.1.2, while the label of the speaker appears in 3.1.1. The primary task during the preprocessing phase involves merging the corpora outlined at 3.1.

A.1 File-level matching

In Hebrew corpus files, there is incompatibility in files' order with English Berman longitudinal dataset. In order to overcome this problem, we made manual changes to the Hebrew corpus, including removing blank files and reordering according to Berman longitudinal dataset files' order.

A.2 Utterance-level matching

In the datasets, most files contain an equal number of lines, except for certain files within the English Berman longitudinal dataset. These additional lines are filled with irrelevant or duplicate information compared to the standard Hebrew data. We manually identified and removed these lines from the English dataset. In this corpus, there are 268 files, out of them 64 are found to be problematic.

A.3 Token-level matching

Matching tokens for each pair of English-Hebrew sentences often leads to numerous conflicts within the sentence (in token level). There are several types of gaps that lead to these conflicts. In the process of overcoming the gaps, we edit the Hebrew sentences in an automatic script.
Here are our primary steps to align as many sentences as feasible, focusing solely on editing Hebrew sentences:

- Create segmented sentences using YAP (Yet Another (natural language) Parser)(More et al., 2019)[12].
- Merge children's names (Hagar, Leor, Lior) to single names instead of separated (for example, Ha gar to Hagar).
- Combine separated words that should be one word.
- Separate conjunction.
- Remove random "junk" letters in the middle of the sentence.
- Insert spaces between punctuation marks that are directly attached to text.
- If punctuation is absent in a Hebrew sentence as it appears in the Latin transcription, add the appropriate punctuation marks.

[12]Apache-2.0 license

- Correct prepositions (if they are written separated in Latin transcription but connected to words in Hebrew).
- Correct double words (combine words like "Od Paam" to "Od_Paam", as in English it appear as a single word - again). We carried out this step at this point rather than earlier because in the preceding sections, we addressed all aspects concerning word indexes when sentences are segmented by spaces.
- Attempt to correct the prepositions once more, considering that the indexes may have changed after addressing duplicate words.

Please note that during the correction of prepositions, we proceeded to the next step only if our function successfully rectified the sentence. If the correction was not made, the incorrect sentence was retained for another attempt.

B Implementation Details and Reproducibility

All experiments were run for 100 epochs, with each run taking approximately 15 minutes of training. For each run, we identified the epoch at which the maximum accuracy on the development set was achieved (referred to as the "max epoch"). The final reported result for each run is the test accuracy at this "max epoch".

During the process, we logged two models in MLflow for each run: the model corresponding to the "max epoch" and the model after completing all 100 epochs. As the top-performing runs showed minimal variation in development and test accuracy, we further refined the process by training the model for each hyper-parameter combination using six different random seeds. The final selected model for each configuration was the one with the highest average development accuracy across these seeds.

B.1 Hyper-parameter optimization

Hyper-parameter optimization was conducted using Optuna (Akiba et al., 2019)[13], an open-source framework designed for efficient and automated hyper-parameter tuning. Optuna employs techniques such as Bayesian optimization and pruning mechanism to enhance search efficiency and reduce computational costs. The optimization process was guided by a defined objective function; maximize the accuracy on the development set and evaluating performance metrics on accuracy and loss.

All experimental results, including hyper-parameter trials, best-performing configurations and model performance metrics, were systematically logged using MLflow[14]. MLflow provided experiment tracking, reproducibility and model versioning, enabling comprehensive monitoring and comparison of different hyper-parameter tuning runs.

B.2 Model Logging

Both the final and best performing models were logged using MLflow. For each of the runs, the model on the last epoch and the best model of the run, selected based on *accuracy_dev_max*, is available for future benchmarking.

C RoBERTa Optimized

In addition to the use of the RoBERTa architecture with the same number of epochs as TafBERTa (see Section 6, we also explore the optimization of the RoBERTa model given the HTBerman data, increasing the number of epochs. The results are presented in Table 6.

D Results Visualization

This appendix provides the detailed evaluation results of various Hebrew language models on grammatical agreement tasks. The models were assessed on Number Agreement, Gender Agreement and Overall Accuracy using the HeCLiMP benchmark. The figures illustrate the performance of each model with respect to the number of parameters and words seen in the training phase.

D.1 Number Agreement Accuracy

The first evaluation metric focuses on the ability of models to correctly predict number agreement in Hebrew. As shown in Figure 2, DictaBERT achieved the highest accuracy at 90.1%, followed by TafBERTa with 80.5%. HeRo performed at 69.1%, while AlephBERT and AlephBERTGimmel recorded 58.7% and 54.3%, respectively. DictaLM 2.0 performed considerably worse than other models with only 31.4%.

[13] MIT License

[14] https://mlflow.org/ with Apache-2.0 license

Hyper-parameter	Checked Intervals
num_attention_heads	{2, 4, 6, 8, 10, 12}
hidden_size	{64, 128, 256, 512, 768}
leave_unmasked_prob	{0.0, 0.1}
num_layers	{2, 4, 6, 8, 10, 12}
intermediate_size	{64, 128, 256, 512, 1024, 2048, 3072, 4096}

Table 5: Intervals checked for each hyperparameter in the Optuna objective function. The upper bound of the search space corresponds to the hyperparameters of RoBERTa. Thus, TafBERTa's smaller size was not intentionally designed to be compact, but rather emerged as the optimal configuration through hyperparameter tuning.

Model	#epoch	Overall	Number	Gender
RoBERTa (HTBerman)	5	65.6	83.7	47.5
RoBERTa (HTBerman)	43	71.1	83.2	59
TafBERTa	5	69.4	80.5	58.2

Table 6: Accuracy on each phenomenon in HeCLiMP using the Holistic-scoring method. "Overall" refers to the overall accuracy across all phenomena. "Number" and "Gender" refer to determiner-noun agreement in number and gender, respectively. RoBERTa was trained for five epochs, matching TafBERTa's training regime and also for a longer period until convergence. When trained for five epochs, RoBERTa achieved lower overall accuracy (**65.6**) compared to TafBERTa (**69.4**), with higher performance on number agreement (**83.7** vs. **80.5**) but weaker results on gender agreement (**47.5** vs. **58.2**). Training RoBERTa for more epochs improved its overall accuracy (**71.1**) and performance on the gender agreement task (**59**) but slightly reduced its accuracy on number agreement (**83.2**). TafBERTa maintains a better balance across both tasks.

D.2 Gender Agreement Accuracy

Figure 3 demonstrates that DictaBERT again performed the best, reaching 84.2 accuracy. HeRo followed with 77.9, while AlephBERTGimmel and AlephBERT obtained 65.3 and 58.5, respectively. TafBERTa recorded 58.2 and DictaLM 2.0 managed 60.8.

D.3 Overall Accuracy

The overall accuracy metric evaluates the general grammatical understanding of Hebrew language models across different agreement phenomena. Figure 4 shows that DictaBERT leads with an 87.1 accuracy, followed by HeRo at 73.5 and TafBERTa at 69.4. AlephBERT and AlephBERTGimmel achieved 58.6 and 59.8, respectively. DictaLM 2.0 recorded an overall accuracy of 46.1, which is notably lower than the other models.

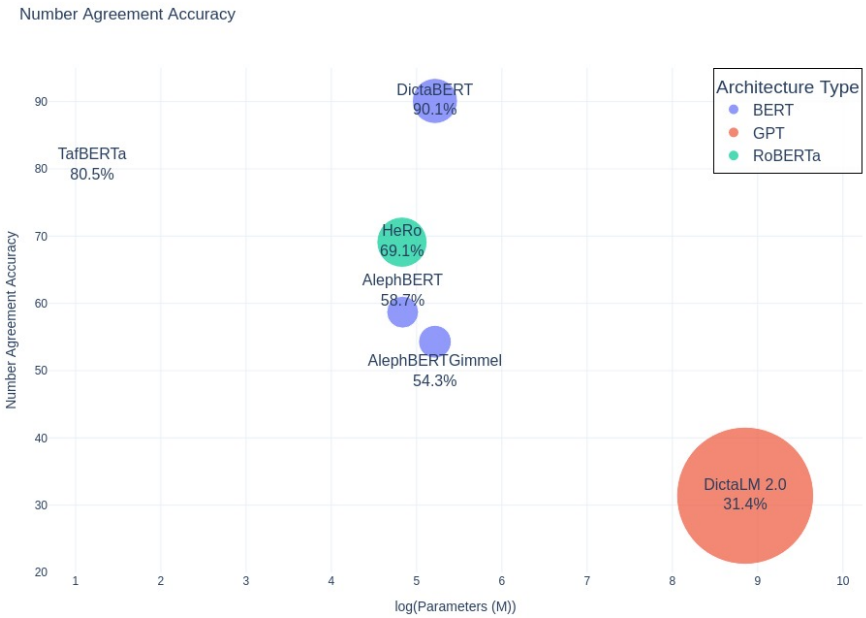

Figure 2: Number Agreement Accuracy of Hebrew Language Models

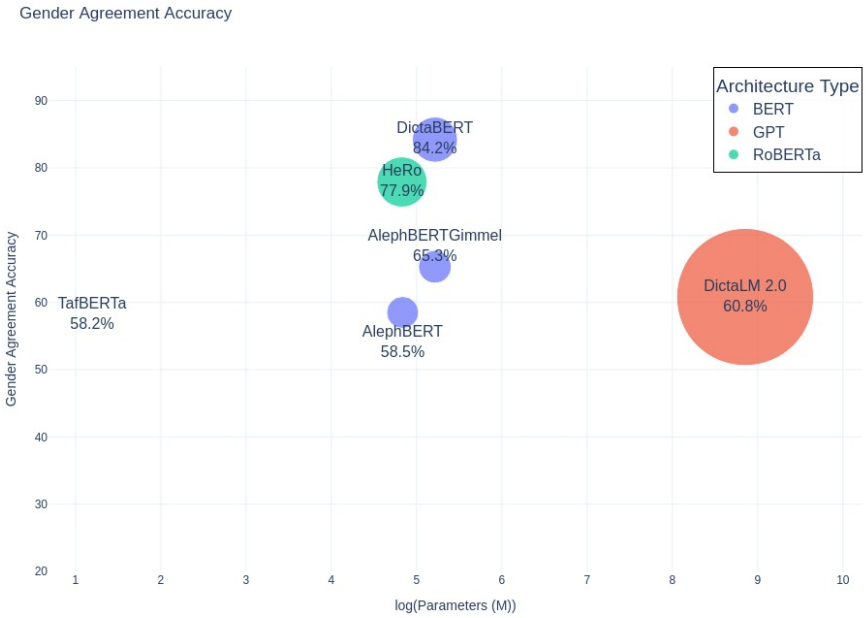

Figure 3: Gender Agreement Accuracy of Hebrew Language Models

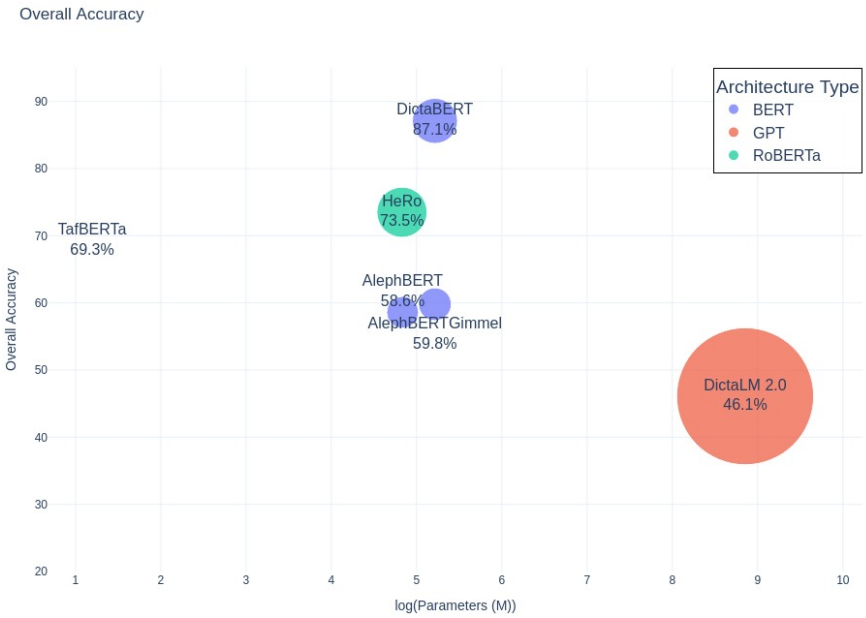

Figure 4: Overall Accuracy of Hebrew Language Models