TafBERTa:
Probing the Grammar Learning Efficiency of Language Models for Hebrew

המחלקה להנדסת מערכות תוכנה ומידע
אוניברסיטת בן-גוריון בנגב
Software and Information Systems Engineering
Ben-Gurion University of the Negev

Ministry of Innovation, Science and Technology
Israel

Anita Gelboim and Elior Sulem

We introduce a **Hebrew version of the BabyBERTa** model, specifically trained on Hebrew **Child Directed Speech**, to evaluate the effectiveness of smaller models in learning Hebrew grammar

1 Motivation

- **Global Motivation**
  - **Green AI**: Minimizes environmental impact with fewer resources
  - **Human-like Learning**: Models should mimic human efficiency, learning language with far less data than LLMs [1]
  - **Democratization**: Make advanced NLP tools accessible to all, not just tech giants
  - **Psycholinguistics**: These models aid in understanding how the brain processes language, supporting research in language acquisition and comprehension
- **Why Hebrew?**
  - **Lagging Behind**: Hebrew language models still fall behind English LLMs
  - **Limited Resources**: Limited data availability in Hebrew compared to English
  - **Morphologically Rich Language**: Hebrew's rich morphology presents unique challenges in handling diverse word forms and structures

וכשמהים

ים	ה	מ	כש	ו
sea	the	from	when	and
NN	DT	IN	REL	CC

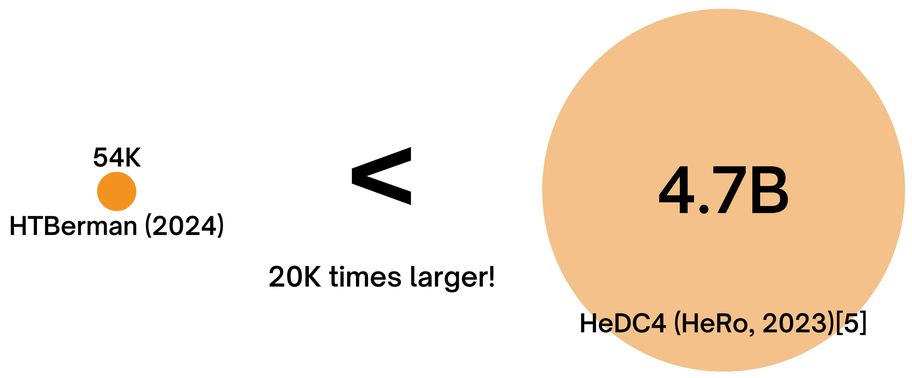
2 Related Work – BabyBerta

A compact variant of the RoBERTa architecture, trained on the CHILDES corpus in **English** on child-directed speech (CDS)

	RoBERTa-base	BabyBERTa [2]	TafBERTa
parameters	125M	8M	3.3M
data size	160GB	0.02GB	1.8MB
words in data	30B	5M	233K
batch size	8K	16	128
max sequence	512	128	128
epochs	>40	10	5
hardware	1024x V100	1x GTX1080	1×RTX6000
training time	24 hours	2 hours	105 seconds
accuracy	81.0	80.5	69.3 (on HeCLiMP)

3 HTBerman - The Dataset [3,4]

Father:	ha#	šafān	ha#	qaṭān	.
%mor:	det|the	n|gen:ms&num:sg	det|the	adj|root:qtn&gen:ms&num:sg	.
	ה	שפן	ה	קטן	.

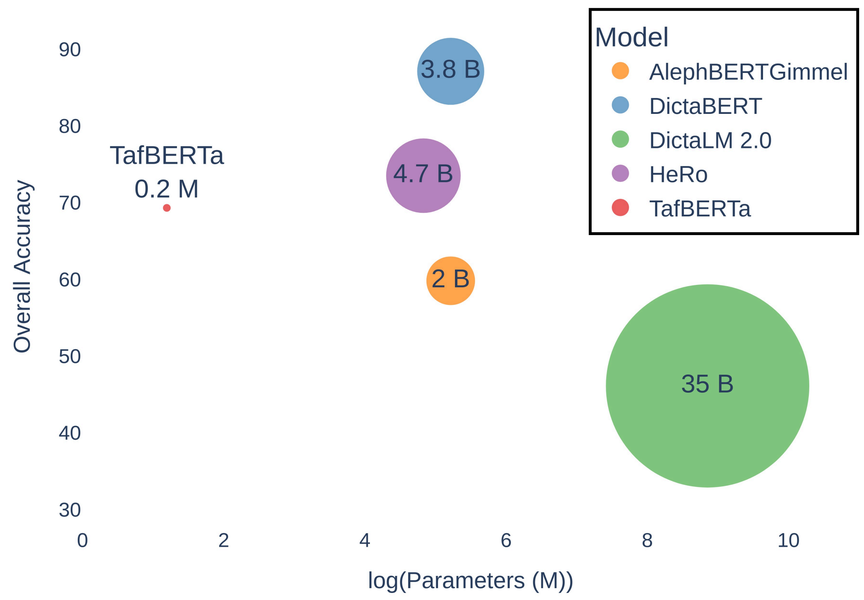


Our model is trained on significantly fewer words than a state-of-the-art (SOTA) model, by several orders of magnitude

4 HeCLiMP - Minimal Pairs Benchmark

Agreement Determiner-Noun Number (Singular-Plural for each gender)	תסתכל על ה כובע הזה . תסתכל על ה כובע האלה .
	תסתכל על ה מחברת הזו . תסתכל על ה מחברת האלה .
Agreement Determiner-Noun Gender (new compared to English)	תסתכל על ה כובע הזה . תסתכל על ה כובע הזו .

5 TafBERTa vs. Other Hebrew Models

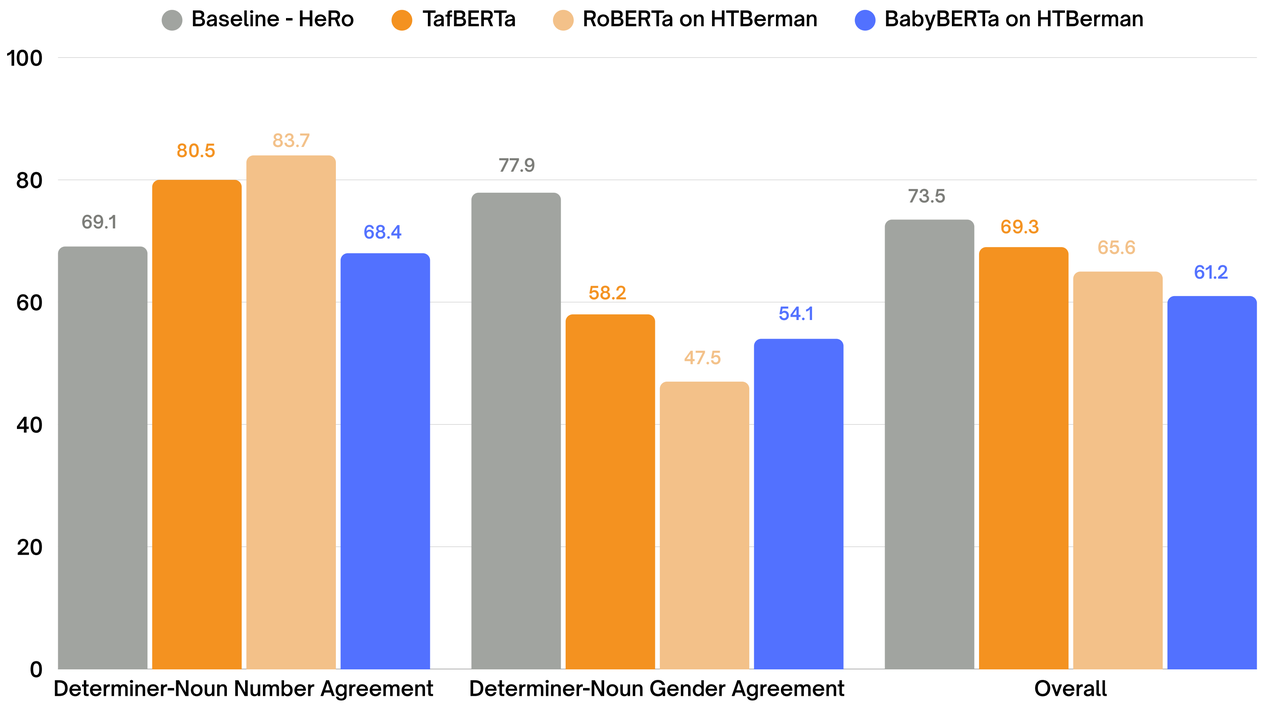


Overall accuracy of Hebrew language models on the HeCLiMP benchmark. Bubble size represents the number of words seen during training, while the x-axis indicates the logarithm of model parameters (M).

Conclusions:

1. **Overall Accuracy** - TafBERTa performs above average
2. **Efficiency vs. Size** - TafBERTa competes well with larger models, proving its efficiency

6 Experiments



Models' accuracy on the grammatical test suite using Holistic score (number of minimal pairs: 3497 for each paradigm)

Conclusions:

1. TafBERTa effectively learns number agreement patterns.
2. Delivers competitive overall performance.
3. Thorough tuning of RoBERTa parameters was essential.
4. Adapting BabyBERTa's architecture to Hebrew was key.

7 Future Work

- **Evaluation Improvements:**
  - Expand HeCLiMP to cover more grammatical structures
- **Multilingual Model Development**
  - Extend TafBERTa to a multilingual framework by training on related Semitic languages (e.g., Arabic)
- Training on Older Children's Data
  - Train on speech to older children, capturing broader linguistic complexity.

Thank you for stopping by!
Let's talk:

Acknowledgment:
Special thanks to Shuly Wintner for the access to the spoken Hebrew CHILDES data :)

References:
[1] What artificial neural networks can tell us about human language acquisition (Warstadt et al., 2024)
[2] BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language (Huebner et al., 2021)
[3] A morphologically-analyzed CHILDES corpus of Hebrew (Nir et al., 2010)
[4] A Morphologically Annotated Hebrew CHILDES Corpus (Albert et al., 2012)
[5] HeRo: RoBERTa and Longformer Hebrew Language Models (Shmidman et al., 2023)