UNIVERSITÄT
BIELEFELD

Fakultät für Linguistik
und Literaturwissenschaft

Linguistic investigations into large and small language models

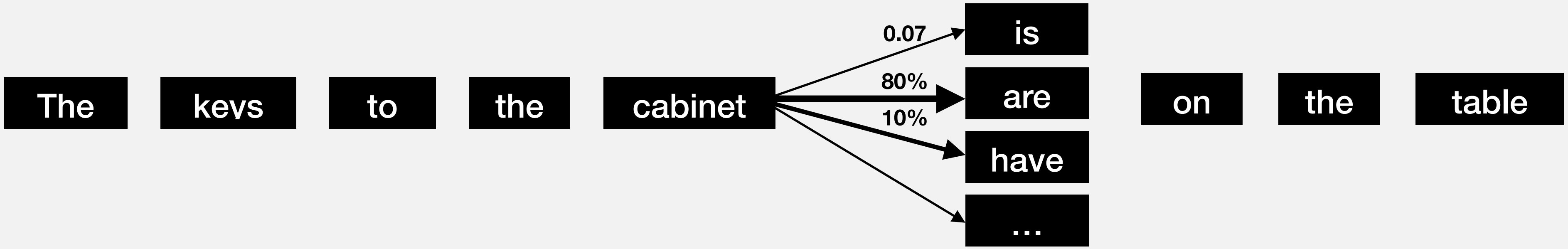
Sina Zarrieß
Bielefeld University

UNIVERSITÄT
BIELEFELD

Fakultät für Linguistik
und Literaturwissenschaft

Large Language Models

- Modern LLMs are trained to predict next tokens given a left context:



- How can such a simple learning architecture capture the complexities of natural language?

UNIVERSITÄT
BIELEFELD

Fakultät für Linguistik
und Literaturwissenschaft

Linguistic investigations into LLMs

- Bertology, GPTology, LODNA (Baroni, 2022): linguistically oriented deepnet analysis
- Does a certain LLM „know“ a certain linguistic rule?

https://direct.mit.edu/coli/article/50/1/293/118131/Language-Model-Behavior-A-Comprehensive-Survey

March 01 2024

Language Model Behavior: A Comprehensive Survey

In Special Collection: CogNet

Tyler A. Chang, Benjamin K. Bergen

Check for updates

Author and Article Information

Computational Linguistics (2024) 50 (1): 293–350.
https://doi.org/10.1162/coli_a_00492 **Article history**

Cite **PDF** **Permissions** **Share**

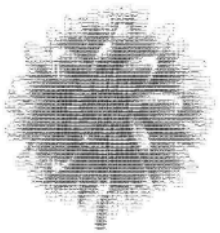
Abstract

Transformer language models have received widespread public attention, yet their generated text is often surprising even to NLP researchers. In this survey, we discuss over 250 recent studies of English language model behavior before task-specific fine-tuning. Language models possess basic capabilities in syntax, semantics, pragmatics, world knowledge, and reasoning, but these capabilities are sensitive to specific inputs and surface features. Despite dramatic increases in generated text quality as models scale to hundreds of billions of parameters, the models are still prone to unfactual responses, commonsense errors, memorized text, and social biases. Many of these weaknesses can be framed as over-generalizations or under-generalizations of learned patterns in text. We synthesize recent results to highlight what is currently known about large language model capabilities, thus providing a resource for applied work and for research in adjacent fields that use language models.

Ongoing linguistic debates around LLMs

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023



By Ruru Kuo

By Noam Chomsky, Ian Roberts and Jeffrey Watumull
Dr. Chomsky and Dr. Roberts are professors of linguistics. Dr. Watumull is a director of artificial intelligence at a science and t
company.

Modern language models refute Chomsky's approach to language

Steven T. Piantadosi[a,b]
[a]UC Berkeley, Psychology [b]Helen Wills Neuroscience Institute

Dissociating language and thought in large language models

Kyle Mahowald[1,5,*], Anna A. Ivanova[2,5,*], Idan A. Blank[3,*], Nancy Kanwisher[4,*], Joshua B. Tenenbaum[4,*], and Evelina Fedorenko[4,*]

Do we need Goliath-style LMs to model language?

- LLMs are strong, but clunky and often easy-to-fool
- It is still not clear what exactly, how and why they learn
- Smaller LMs let us do smarter experiments: data manipulation, deeper analyses, model variations

UNIVERSITÄT BIELEFELD
Fakultät für Linguistik und Literaturwissenschaft

LINCC SFB 1646

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How Small Can a Small Language Model Really Get?. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Bastian Bunzeck and Sina Zarrieß. 2024. Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly. In *Proceedings of MILLing 2024, Gothenburg*.

Bastian Bunzeck, Daniel Duran, Leonie Schade and Sina Zarrieß. 2024. Small Language Models Like Small Vocabularies: Probing the Linguistic Abilities of Grapheme- and Phoneme-Based Baby Llamas. https://arxiv.org/pdf/2410.01487

UNIVERSITÄT
BIELEFELD

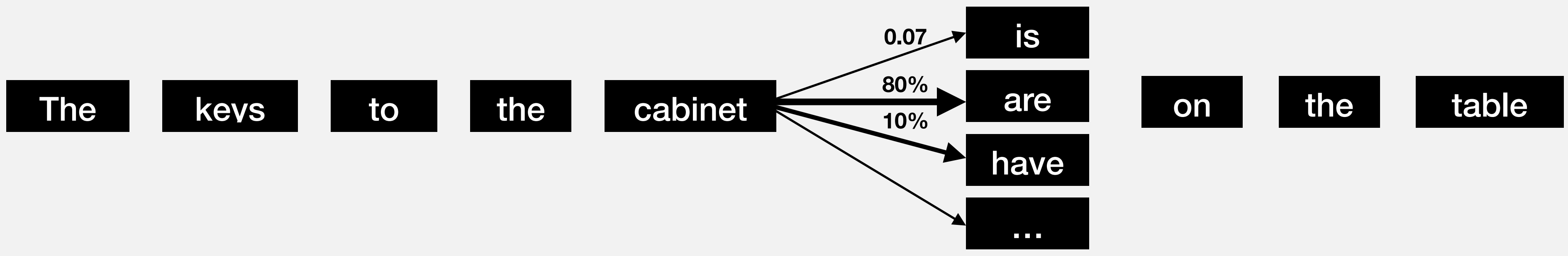
Fakultät für Linguistik
und Literaturwissenschaft

Outline

- **Syntactic knowledge in large and small LMs**
- Syntactic learning trajectories in medium-to-small LMs
- Lexical and syntactic learning in small LMs
- Current directions

Why syntax?

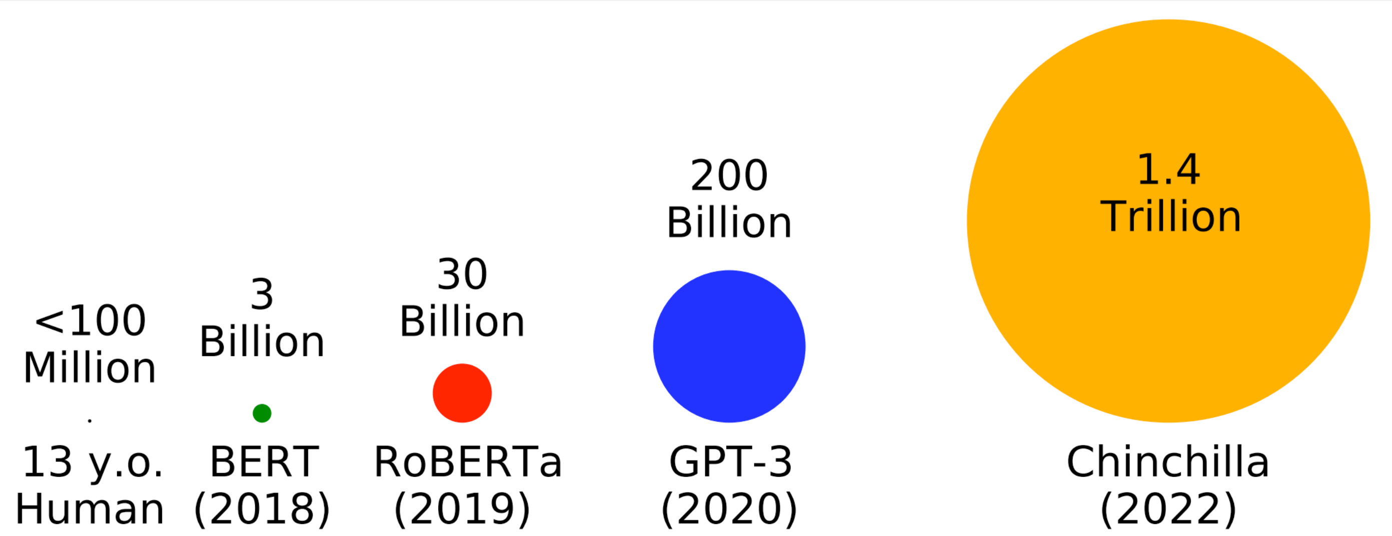
- LLMs are trained on linear sequences of tokens:



- Do LMs learn/represent hierarchical structures in language from linear next token prediction?

What data should LMs be trained on?

- LLM training data massively exceeds human input
- Do we need massive data to learn language with a small LM?



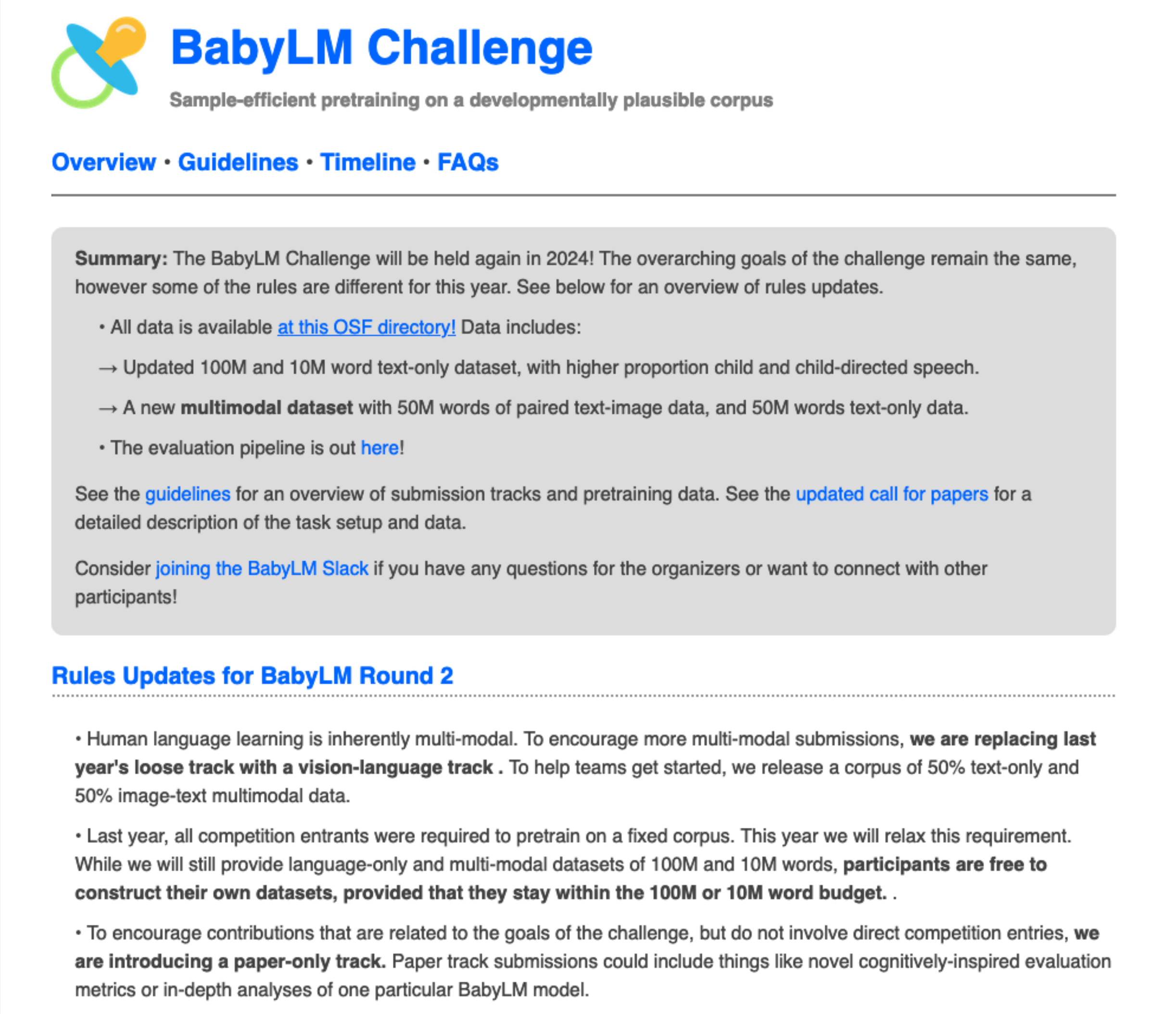
https://babylm.github.io/

UNIVERSITÄT BIELEFELD

Fakultät für Linguistik und Literaturwissenschaft

The BabyLM challenge

- Shared task at CoNLL 2023, and again in 2024
- Task: LM pretraining on a 100M or 10M dataset
- Evaluation: BLiMP (+ SuperGLUE, Age-of-Aquisition prediction)



BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus

Overview · Guidelines · Timeline · FAQs

Summary: The BabyLM Challenge will be held again in 2024! The overarching goals of the challenge remain the same, however some of the rules are different for this year. See below for an overview of rules updates.

- All data is available at this OSF directory! Data includes:

→ Updated 100M and 10M word text-only dataset, with higher proportion child and child-directed speech.

→ A new **multimodal dataset** with 50M words of paired text-image data, and 50M words text-only data.

- The evaluation pipeline is out here!

See the guidelines for an overview of submission tracks and pretraining data. See the updated call for papers for a detailed description of the task setup and data.

Consider joining the BabyLM Slack if you have any questions for the organizers or want to connect with other participants!

Rules Updates for BabyLM Round 2

- Human language learning is inherently multi-modal. To encourage more multi-modal submissions, **we are replacing last year's loose track with a vision-language track .** To help teams get started, we release a corpus of 50% text-only and 50% image-text multimodal data.
- Last year, all competition entrants were required to pretrain on a fixed corpus. This year we will relax this requirement. While we will still provide language-only and multi-modal datasets of 100M and 10M words, **participants are free to construct their own datasets, provided that they stay within the 100M or 10M word budget.** .
- To encourage contributions that are related to the goals of the challenge, but do not involve direct competition entries, **we are introducing a paper-only track.** Paper track submissions could include things like novel cognitively-inspired evaluation metrics or in-depth analyses of one particular BabyLM model.

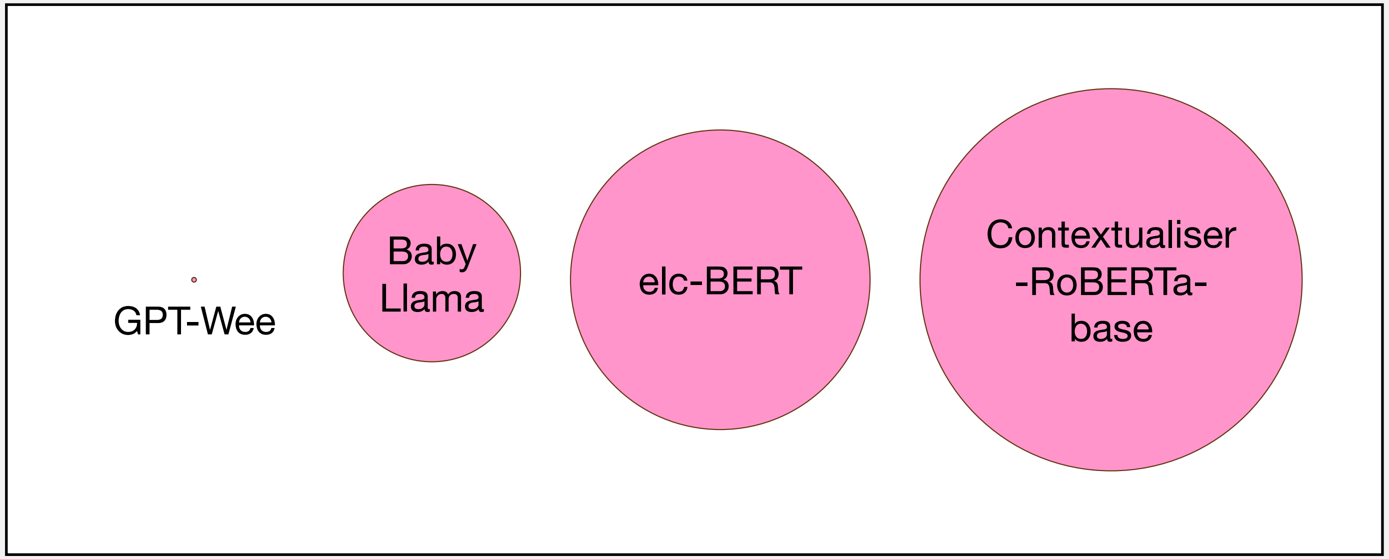
BLiMP - The Benchmark of Linguistic Minimal Pairs for English
(Warstadt et al. 2020)

- Subject-verb agreement:
  - *The sisters bake.* vs. **The sisters bakes.*
- Irregular froms:
  - *Aaron broke the bike.* vs. **Aaron broken the bike.*
- Causatives:
  - *Aaron breaks the glass*. vs. **Aaron appeared the glass.*

- **Accuracy-based evaluation:** Does the LM assign higher probs to the grammatical sentence?

GPT-Wee: How small can a BabyLM be?
(Bunzeck and Zarrieß, 2023)

- Our model @BabyLM 2023:
  - 1.55M parameters
  - Rank 104/121 submissions for strict-small track
  - One of the smallest models submitted (maybe actually the smallest!)
  - Generative architecture



BLiMPing GPT-wee

- GPT-wee performance is decent on all tasks (rarely worse than the worst LLM baseline)
- GPT-wee matches or exceeds LLM performance on some tasks: filler gap, irregular forms, …

	anaphor agreement	argument structure	binding	control raising	determiner noun agreement	ellipsis	filler gap	irregular forms	island effects
16k	73.82	71.91	68.97	66.26	88.36	54.56	68.67	86.06	41.03
16k (cu.)	82.87	69.51	65.24	63.21	85.52	55.43	66.65	77.56	40.88
OPT	63.8	70.6	67.1	66.5	78.5	62	63.8	67.5	48.6
RoBERTa	81.5	67.1	67.3	67.9	90.8	76.4	63.5	87.4	39.9
T5	68.9	63.8	60.4	60.9	72.2	34.4	48.2	77.6	45.6

Overview of BabyLM architectures

- Encoders outperform decoders
- Among the decoders, BabyLlama performs best

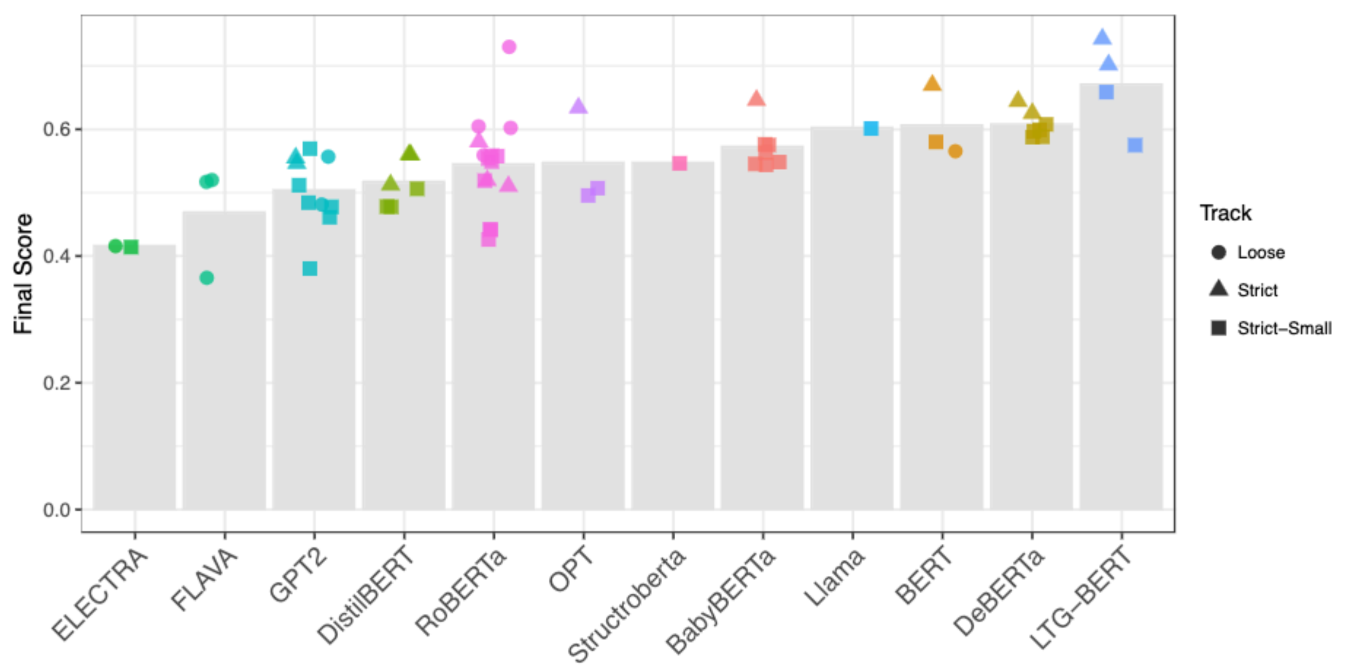


Figure 6: **Effect of Backbone Architecture:** Each point represents a submission. Shape indicates the challenge track. Gray bars show within-category aggregates.

https://aclanthology.org/2023.conll-babylm.1.pdf

Upcoming: 50 shapes of BLiMP
(Bunzeck and Zarrieß, 2024, MiLLing)

- Syntax is learned from much smaller amounts of data than training data of LLMs
- We still do not understand the relationship between model size, data size, and syntactic knowledge in LMs

	Param.	Train. tokens	Hddn. layers	Attn. heads	Embed. size	BLiMP score
baby_llama	2.97M	10M	8	8	128	64%
teenie_llama	2.97M	100M	8	8	128	67%
weenie_llama	11.44M	10M	16	16	256	67%
tweenie_llama	11.44M	100M	16	16	256	71%
pythia-14m	14M	300B	6	4	512	65%
pythia-70m	70M	300B	6	8	512	75%
pythia-160m	160M	300B	12	12	768	79%
pythia-410m	410M	300B	24	16	1024	82%
pythia-1b	1B	300B	16	8	2048	82%
pythia-1.4b	1.4B	300B	24	16	2048	82%

Table 1: Model hyperparameters of our self-trained llama models and the compared pythia models

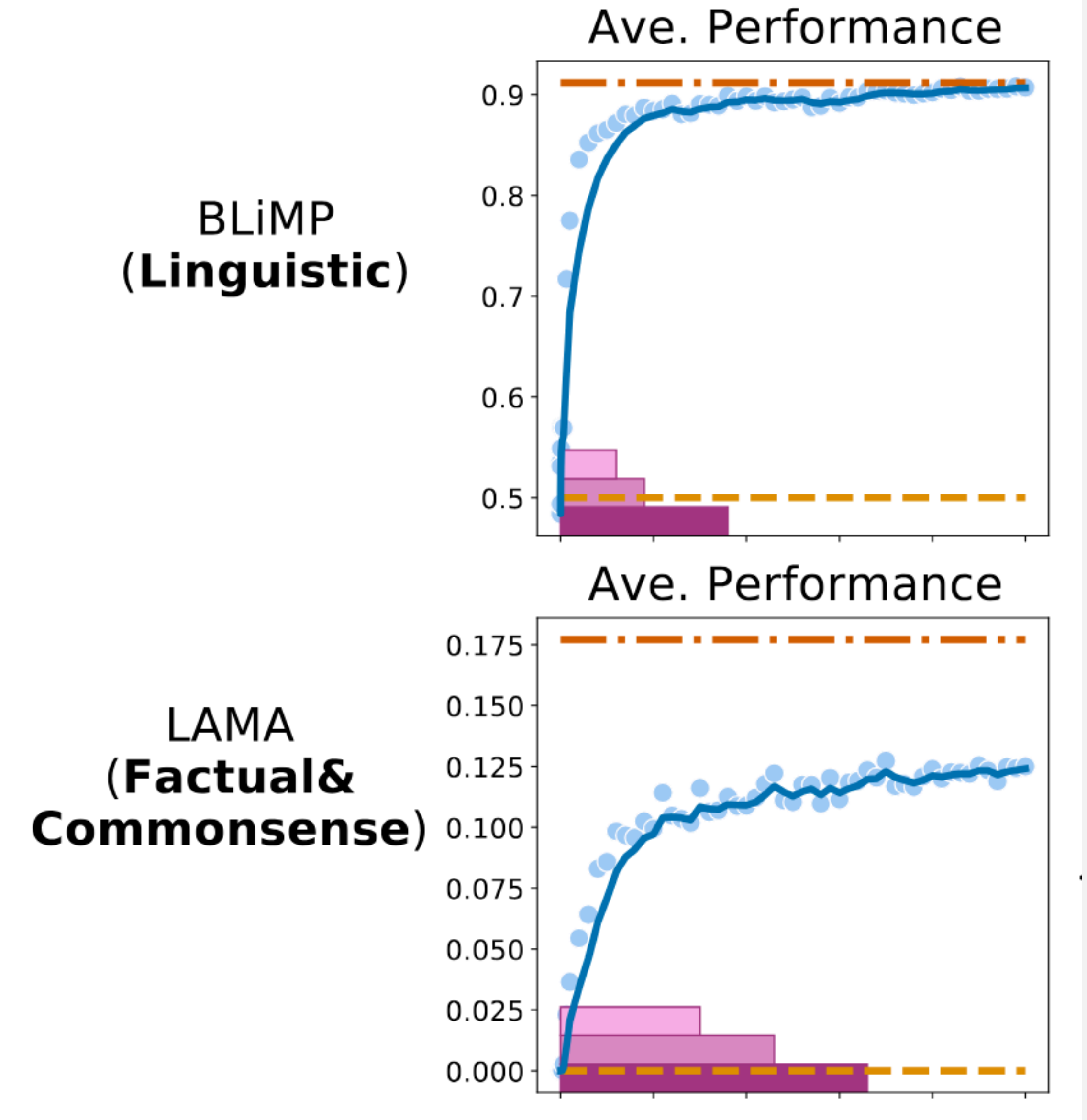
Outline

- Syntactic knowledge in large and small LMs
- **Syntactic learning trajectories in medium-to-small LMs**
- Lexical and syntactic learning in small LMs
- Current directions

UNIVERSITÄT BIELEFELD
Fakultät für Linguistik und Literaturwissenschaft

Learning trajectories

- How does the performance of LMs on linguistic benchmarks develop over the training process?
- Liu et al (2021) probe RoBERTa across time: syntax learning is really fast and stable
- But: Recent LLMs mostly do not provide fine-grained checkpoints

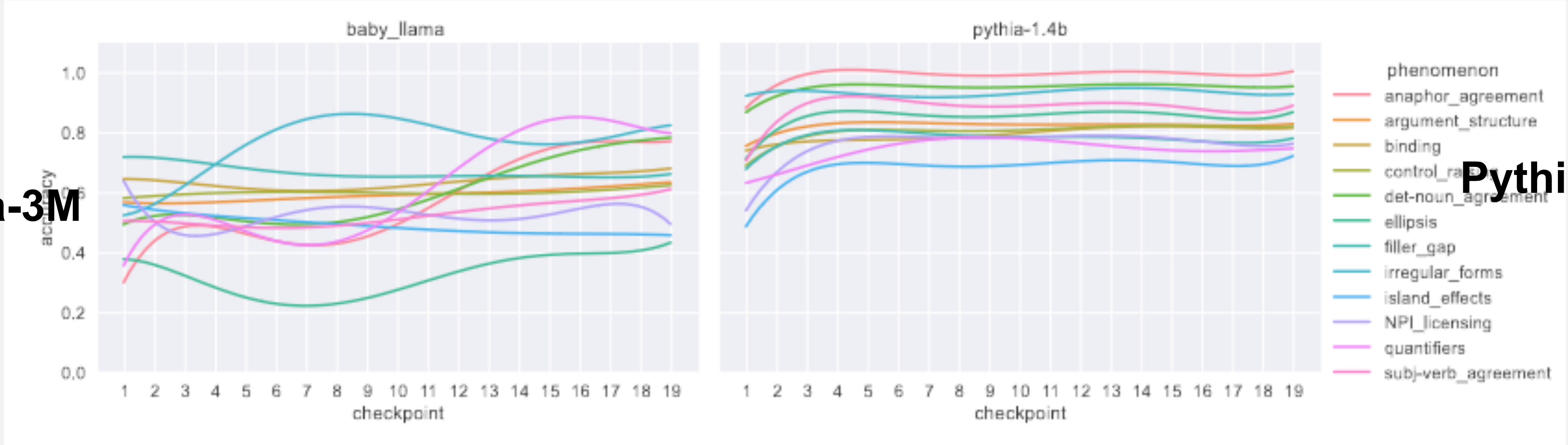


https://aclanthology.org/2021.findings-emnlp.71.pdf

UNIVERSITÄT
BIELEFELD
Fakultät für Linguistik
und Literaturwissenschaft

Is syntax learning really so early and stable in LMs?
(Bunzeck and Zarrieß, 2024, MiLLing)

BabyLlama-3M



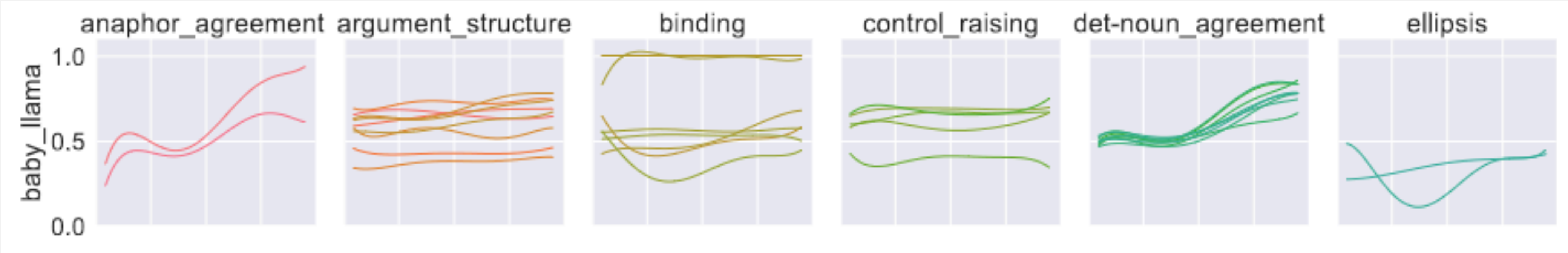
Pythia-1.4B

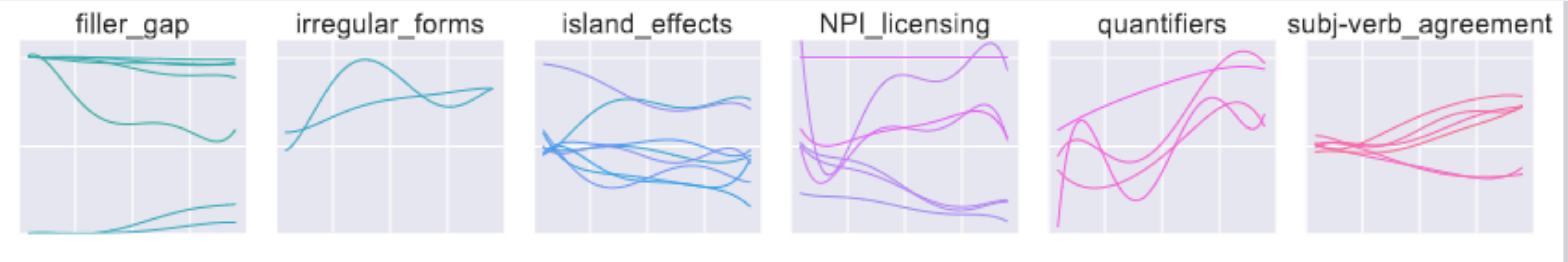
- Checkpoints: 1st training epoch of our baby-llama (10M words) and pythia models (300B words)
- Curves: averaged over phenomena within a BLiMP paradigm (agreement, binding, filler gap, etc.)

UNIVERSITÄT
BIELEFELD
Fakultät für Linguistik
und Literaturwissenschaft

Zooming into BabyLlama’s syntax learning

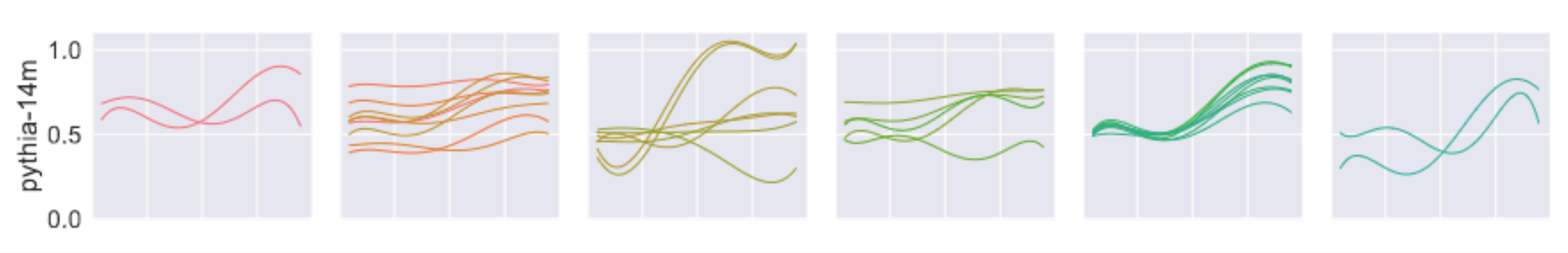
- Individual phenomena in BLiMP are learned with different trajectories
- Shapes: flat, exponential, s-shaped, u-shaped, ill-behaved

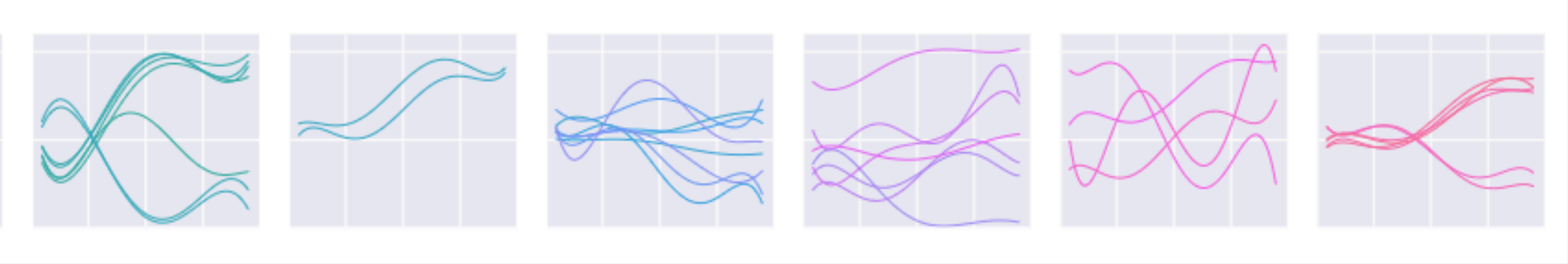




Zooming into Pythia’s syntax learning

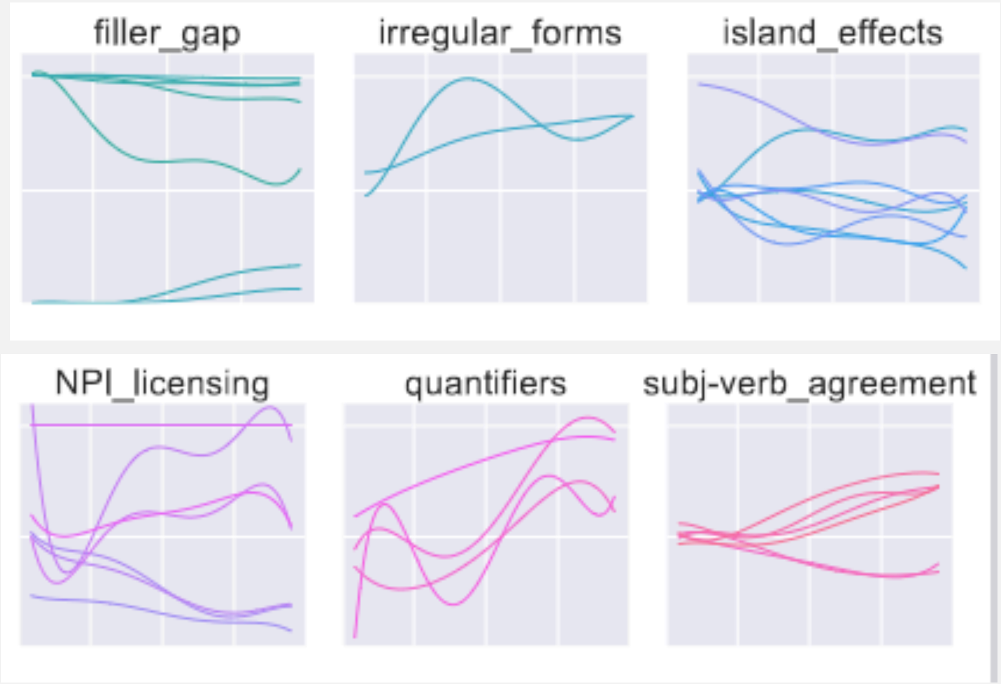
- We find the same range of shapes in the bigger Pythia models
- Shapes in big and small models are often similar for the same phenomenon





Smaller models, more insights

- Shapes of learning curves for individual phenomena vary
- Ill-behaved curves occur (also in bigger models!, around 25% of BLiMP test sets)
- Some paradigms in BLiMP show with very consistent shapes of curves
- We observe „turning points“ within and across paradigms where curves go up for X and down on Y
- But: more work is needed to „classify“ trajectories



UNIVERSITÄT
BIELEFELD
Fakultät für Linguistik
und Literaturwissenschaft

Outline

- Syntactic knowledge in large and small LMs
- Syntactic learning trajectories in medium-to-small LMs
- **Lexical and syntactic learning in small LMs**
- Current directions

Lexical and phonological learning in small LMs

- Learning below the syntax level (morphology, phonology) is completely understudied in LLMs
- Most LLMs come with (sub)word-level tokenization
- Can small models learn word- and syntax-level knowledge?

Small Language Models Like Small Vocabularies: Probing the Linguistic Abilities of Grapheme- and Phoneme-Based Baby Llamas

Bastian Bunzeck, Daniel Duran, Leonie Schade and **Sina Zarrieß**
Department of Linguistics
Bielefeld University, Germany
firstname.lastname@uni-bielefeld.de

Abstract

Current language models use subword-based tokenization algorithms like Byte Pair Encoding, which put their validity as models of linguistic representations into question. In this paper, we explore the potential of tokenization-free, phoneme- and grapheme-based language models. We demonstrate that small models based on the Llama architecture can achieve strong linguistic performance on standard syntactic and novel lexical/phonetic benchmarks when trained with character-level vocabularies. We further show that phoneme-based models without any graphemic biases almost match grapheme-based models in standard tasks and novel evaluations. Our findings suggest a promising direction for creating more linguistically plausible language models that are better suited for computational studies of language acquisition and processing.

In this paper, we train and evaluate small Llama models (Touvron et al., 2023) on input that is not pre-segmented into words. Instead, we treat the individual characters in our training data as tokens, meaning that the LM does not receive any prior information on what "meaningful" units in the input are. We investigate whether these small models trained with drastically smaller, linguistically more plausible vocabularies still achieve comparable performance on evaluations across different linguistic levels, i.e. syntax, lexicon and phonetics. Additionally, we compare models trained on graphemes and models trained on phonemes[1], questioning the common assumption that grapheme-based learners are as *tabula rasa* (Hahn and Baroni, 2019) as LMs can get.

We find that our character-based LMs perform as well on standard evaluation measures as comparable subword-based models trained on the same

https://arxiv.org/pdf/2410.01487

Benchmarking small grapheme and phoneme LMs

- We train on BabyLM data
- We convert text to phoneme sequences with G2P
- We generate non-words with `wuggy`, and test for lexical decision performance

Example (graphemic)	Example (phonetic)
BLiMP (Minimal pairs)	
👍 Aaron breaks the glass.	👍 ɛɹʌn bɹeɪks ðʌ glæs
👎 Aaron appeared the glass.	👎 ɛɹʌn ʌpɪɹd ðʌ glæs
Lexical decision task (Minimal pairs)	
👍 drunk.	👍 drʌŋk
👎 blunk.	👎 frʌŋk
Rhyme prediction (Probing)	
✔️ The sky was clear, but full of cheer.	✔️ ðʌ skaɪ wɑz klɪɹ bʌt fʊl ʌv tʃɪɹ
❌ The door opened with a creak.	❌ ðʌ dɔɹ oʊpʌnd wɪð ʌ kɹik
Age prediction (Probing)	
👶 rock , rock , rock .	👶 waːwaːwaː
🧒 hold my juice Mommy .	🧒 hod mai ʤus mami
👦 open the door .	👦 opən ðə dər

Table 1: Examples of all evaluation paradigms

Results

Evaluation	Grapheme model	Grapheme model, no whitesp.	Phoneme model	Phoneme model, no whitesp.	BabyLlama
BLiMP	71.69%	68.88%	66.90%	64.88%	73.10%
BLiMP supplement	52.30%	56.28%	55.42%	54.13%	60.60%
Lexical decision task	99.00%	99.10%	68.20%	63.80%	69.00%
Rhyme prediction	88.50%	91.50%	85.00%	78.49%	92.50%
Age prediction	60.50%	58.90%	61.10%	57.80%	60.90%

Table 2: Evaluation results: for BLiMP and the lexical decision task, the scores correspond to the percentage of correct choices in a minimal pair setting; for rhyme and age prediction the scores report classification accuracy.

- Grapheme LM outperforms BabyLlama on lexical decision
- Grapheme LM is close to BabyLlama on BLiMP
- Phoneme LMs perform slightly worse

Outline

- Syntactic knowledge in large and small LMs
- Syntactic learning trajectories in medium-to-small LMs
- Lexical and syntactic learning in small LMs
- **Current directions**

Compiling a BabyLM corpus for German

- The English BabyLM corpus:

Dataset	Domain	# Words		Proportion
		Strict-Small	*Strict*	
CHILDES (MacWhinney, 2000)	Child-directed speech	0.44M	4.21M	5%
British National Corpus (BNC),[1] dialogue portion	Dialogue	0.86M	8.16M	8%
Children's Book Test (Hill et al., 2016)	Children's books	0.57M	5.55M	6%
Children's Stories Text Corpus[2]	Children's books	0.34M	3.22M	3%
Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2020)	Written English	0.99M	9.46M	10%
OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	3.09M	31.28M	31%
QCRI Educational Domain Corpus (QED; Abdelali et al., 2014)	Educational video subtitles	1.04M	10.24M	11%
Wikipedia[3]	Wikipedia (English)	0.99M	10.08M	10%
Simple Wikipedia[4]	Wikipedia (Simple English)	1.52M	14.66M	15%
Switchboard Dialog Act Corpus (Stolcke et al., 2000)	Dialogue	0.12M	1.18M	1%
Total	–	9.96M	98.04M	100%

Table 1: The datasets we release for the *Strict* and *Strict-Small* tracks of the BabyLM Challenge. We present the number of words in the training set of each corpus that we include. [1]http://www.natcorp.ox.ac.uk [2]https://www.kaggle.com/datasets/edenbd/children-stories-text-corpus [3]https://dumps.wikimedia.org/enwiki/20221220/ [4]https://dumps.wikimedia.org/simplewiki/20221201/

Analyzing Pragmatics in Small LMs

- **Judith Sieker** and Sina Zarrieß. 2023. When Your Language Model Cannot Even Do Determiners Right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle. In *Proceedings of the 6th BlackboxNLP Works*

Context: Jan's mother was shopping. She bought one banana and two pears.

(a) **Unique fruit**: Of these, Jan received [the | a] banana.

(b) **Non-unique fruit**: Of these, Jan received [a | the] pear.

(c) **Pair of fruits**: Of these, Jan received [both | all] pears.

Figure 1: Exemplified conditions of our study (images included for illustration purposes only).

Analyzing linguistic creativity with LMs

- Testing the dual-route account of phonological encoding with LMs
- Training on conversational, spoken data

A02: Creating novel phonetic representations across varying communication settings

PIs: Prof. Dr. Joana Cholin/ Prof. Dr. Petra Wagner/ Prof. Dr. Sina Zarrieß

In speech, deviations from canonical realisations of phonemes, syllables or larger units are very common. A02 aims to under-stand the creative flexibility of the processes involved in such productions via experimental production studies and psycholin-guistic and computational modelling. We will investigate whe-ther and how creatively constructed phonetic forms can be selectively elicited and modelled in different interactive and lin-

Summary

- Model size does not seem to be the key for learning ``core'' linguistic knowledge

- Much more systematic experimentation is needed:
  - which training data benefits learning?
  - which architectural decisions matter?
  - where is the sweet spot between general performance and modeling flexibility?

- These can be easily done with smart little BabyLMs!

UNIVERSITÄT
BIELEFELD
Fakultät für Linguistik
und Literaturwissenschaft

Thank you!

LINCC
SFB 1646

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How Small Can a Small Language Model Really Get?. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Bastian Bunzeck and Sina Zarrieß. 2024. Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly. In *Proceedings of MILLing 2024, Gothenburg*.

Bastian Bunzeck, Daniel Duran, Leonie Schade and Sina Zarrieß. 2024. Small Language Models Like Small Vocabularies: Probing the Linguistic Abilities of Grapheme- and Phoneme-Based Baby Llamas. https://arxiv.org/pdf/2410.01487