UNIVERSITY
OF OSLO

BabyLM Challenge: What is it? Best models and Papers

Lucas Georges Gabriel Charpentier

lgcharpe@ifi.uio.no

Language Technology Group, University of Oslo

14th December 2023

Contents

Introduction

- Challenge proposed by Alex Warstadt et al.
- Creation of a small but high-quality dataset to match the number of tokens a 13-year-old child is exposed to.
- Plan to have multiple iterations of the challenge.

Motivation

1. Creating more cognitively plausible models.
2. Optimizing training pipelines before scaling.
3. Democratizing language model pre-training outside industry.

Tracks

- 3 tracks proposed:
  1. STRICT
  2. STRICT-SMALL
  3. LOOSE

STRICT Track

- Combination of 10 different datasets including:
  1. Developmentally plausible domains (child-directed speech, transcribed dialogue, and children’s literature)
  2. Encyclopedic knowledge (Wikipedia and Wikipedia simple)
  3. Complex written English (Guttenberg project)
  4. Subtitles (movie and educational videos)
- The dataset contains 100M words.
- Only models trained with this dataset can be used.

STRICT-SMALL Track

- A scaled down to 10M word version of the STRICT track dataset.
- As for the previous track, only models trained on this dataset can be submitted.

LOOSE Track

- Any language data possible but with a limit of 100M words total.
- Unlimited use of other data types (audio, image, etc.).
- Enabled to the possibility of multimodality.

BLiMP

- Used to evaluate the grammatical abilities of LMs.
- A minimal pair of sentences, one acceptable and the other not.
- If the model assigns a higher probability to the acceptable sentence then it is correct.

BLiMP Supplemental

- Same evaluation style as BLiMP.
- 5 additional tasks: Hypernyms, Subject-auxiliary inversion, turn-talking, question-answer congruence (easy+tricky)
- Tests the model’s linguistic knowledge of questions and dialogue.

(Super)GLUE

- Mix of the GLUE and SuperGLUE benchmarks.
- Test LMs ability on downstream tasks (mainly text classification tasks).
- Includes:
  1. Paraphrase detection (MRPC, QQP)
  2. Sentiment classification (SST-2)
  3. Natural Language Inference (MNLI, QNLI, RTE)
  4. Question-answering (BoolQ, MultiRC)
  5. Acceptability judgements (CoLA)
  6. Commonsense Reasoning (WSC)

MSGS

- Tests whether models bias linguistic or surface features.
- Trained on ambiguous data, containing both feature types or neither.
- Evaluated on unambiguous data with labels indicating the presence of the linguistic feature.
- Score of -1 = Surface bias, 1 = Linguistic bias
- Surface features include lexical content and relative token position.
- Linguistic features include main verb form, syntactic category, and control raising.

Findings

- **Helpful**: Knowledge distillation from auxiliary models and data preprocessing
- **Mixed/Unclear**: Curriculum learning and model scaling
- **Not helpful**: Multimodal learning and training objectives

BabyLM Challenge 2024

- BabyLM 2024 is confirmed.
- Deadlines and conference TBA.
- Potential changes:
    - Focus on multimodal, i.e. more loose tracks.
    - Limitations on training epochs/steps/flops.
    - Standardized pipelines to preprocess data.

Survey

- Survey on the BabyLM challenge available at `https://babylm.github.io/`
- Fill it in if you have ideas, or suggestions for the next iterations.

Contents

Introduction

- **Motivation**: Standard transformer-based models use standard residuals that weigh all layers equally.
- **Goal**: See whether learning layer weights produce different weighing for each layer while retaining performance.
- **Constraints**: Using a small (100M and 10M words) but good quality dataset to pre-train the models.

LTG BERT

- For all other training choices, we adapt the approach of LTG-BERT.
- This model was optimized for low-resource MLM on a similar corpus.
- LTG-BERT uses several improvements:
  1. NormFormer layer normalization,
  2. a disentangled attention mechanism with relative positions (DeBERTa),
  3. GEGLU activation function,
  4. high weight decay,
  5. no linear biases,
  6. random span masking

BabyLM Datasets

1. STRICT track:
   - Used to train base version (∼100M parameters) of LTG-BERT and ELC-BERT
2. STRICT-SMALL track:
   - Used to train small version (∼25M parameters) of LTG-BERT and ELC-BERT

Preprocessing

- The pretraining datasets for the STRICT and STRICT-SMALL tracks are a mix 10 different corpora.
- We applied light preprocessing and normalization to these corpora to convert them into a unified format
- For example, in the CHILDES subcorpus, the preprocessing:
  1. capitalizes the first letter of each line,
  2. normalizes punctuation and whitespaces (detokenization),
  3. puts every line between double quotes (as directed speech).

Preprocessing

- Similar steps are done for other subcorpora and in addition:
  - We replace some remnants of the Penn Tree format in Children's Book Test (`-LRB-` and `-RRB-` tokens are replaced by '(' and ')'),
  - We restore the original paragraphs of Project Gutenberg (the text file is aligned into blocks by inserting a newline symbol after at most 70 characters, which ruins the sentence structure)

UNIVERSITY
OF OSLO

Residuals

Original residual connection:

$$\boldsymbol{h}_{\text{in}}^{n} \leftarrow \boldsymbol{h}_{\text{out}}^{n-1} + \boldsymbol{h}_{\text{in}}^{n-1}$$

Standard encoder flow:

$$\begin{aligned}
\boldsymbol{h}_{\text{out}}^{0} &\leftarrow \text{embedding}(\boldsymbol{x}), \\
\boldsymbol{h}_{\text{out}}^{n} &\leftarrow \text{att}(\boldsymbol{h}_{\text{in}}^{n}) + \text{mlp}\big(\boldsymbol{h}_{\text{in}}^{n} + \text{att}(\boldsymbol{h}_{\text{in}}^{n})\big)\,, \\
\boldsymbol{y} &\leftarrow \text{LM_head}\big(\textstyle\sum_{i=0}^{N} \boldsymbol{h}_{\text{out}}^{i}\big)
\end{aligned}$$

Modifications

New residual connection:

$$\boldsymbol{h}_{\text{in}}^{n} \leftarrow \sum_{i=0}^{n-1} \alpha_{i,n} \boldsymbol{h}_{\text{out}}^{i}$$

New encoder flow:

$$\begin{aligned}
\boldsymbol{h}_{\text{out}}^{0} &\leftarrow \text{embedding}(\boldsymbol{x}), \\
\boldsymbol{h}_{\text{out}}^{n} &\leftarrow \text{att}(\boldsymbol{h}_{\text{in}}^{n}) + \text{mlp}\left(\text{att}(\boldsymbol{h}_{\text{in}}^{n})\right), \\
\boldsymbol{y} &\leftarrow \text{LM_head}(\boldsymbol{h}_{\text{out}}^{N})
\end{aligned}$$

Ablation Modifications

1. Adding the internal residual:

$$\boldsymbol{h}^n_{\text{out}} \leftarrow \text{att}(\boldsymbol{h}^n_{\text{in}}) + \text{mlp}\big(\boldsymbol{h}^n_{\text{in}} + \text{att}(\boldsymbol{h}^n_{\text{in}})\big)$$

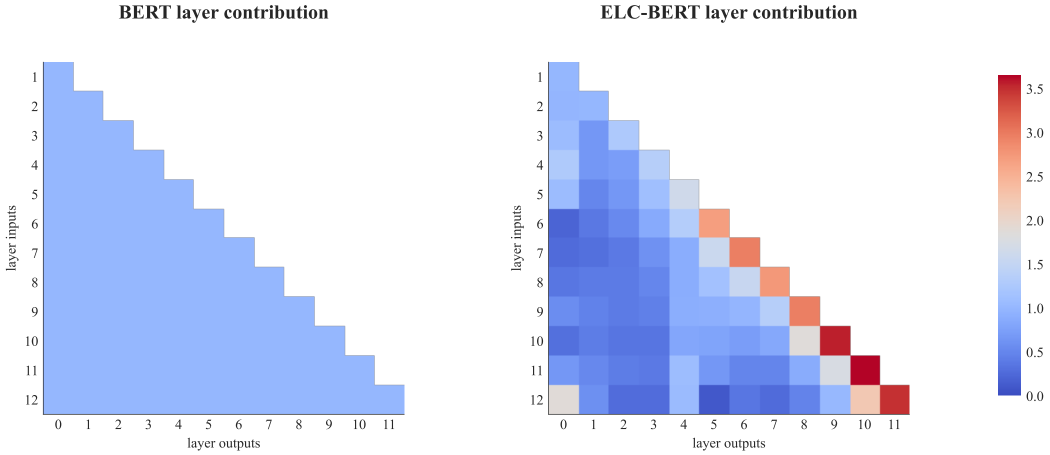
2. Zero initialization: we initialize all the α as equal.
3. Normalization: We add the following step to our encoder layer:

$$\boldsymbol{h}^n_{\text{out}} \leftarrow \text{LayerNorm}(\boldsymbol{h}^n_{\text{out}})$$

4. Weighted output:

$$\boldsymbol{y} \leftarrow \text{LM\_head}\left(\sum_{i=0}^{N} \alpha_{i,o} \boldsymbol{h}^i_{\text{out}}\right)$$

Layer Weighting



Base Model Results

STRICT-SMALL track (10M words)

Model	BLiMP	Supp.	MSGS	GLUE
OPT_{125m}	62.6	54.7	$-0.64^{\pm 0.1}$	$68.3^{\pm 3.3}$
RoBERTa_{base}	69.5	47.5	$-0.67^{\pm 0.1}$	$72.2^{\pm 1.9}$
T5_{base}	58.8	43.9	$-0.68^{\pm 0.1}$	$64.7^{\pm 1.3}$
LTG-BERT_{small}	**80.6**	**69.8**	$\mathbf{-0.43}^{\pm 0.4}$	$74.5^{\pm 1.5}$
ELC-BERT_{small}	80.5	67.9	$-0.45^{\pm 0.2}$	$\mathbf{75.3}^{\pm 2.1}$

STRICT track (100M words)

Model	BLiMP	Supp.	MSGS	GLUE
OPT_{125m}	75.3	67.8	$-0.44^{\pm 0.1}$	$73.0^{\pm 3.9}$
RoBERTa_{base}	75.1	42.4	$-0.66^{\pm 0.3}$	$74.3^{\pm 0.6}$
T5_{base}	56.0	48.0	$-0.57^{\pm 0.1}$	$75.3^{\pm 1.1}$
LTG-BERT_{base}	**85.8**	**76.8**	$-0.42^{\pm 0.2}$	$77.9^{\pm 1.1}$
ELC-BERT_{base}	85.3	76.6	$\mathbf{-0.26}^{\pm 0.5}$	$\mathbf{78.3}^{\pm 3.2}$

Ablations Results

Model	BLiMP	Supp.	MSGS	GLUE
ELC-BERT	85.3	76.6	$-0.26^{\pm 0.5}$	$78.3^{\pm 3.2}$
+ zero initialization	84.9	**78.5**	$-0.38^{\pm 0.3}$	$\mathbf{79.4}^{\pm 1.0}$
+ normalization	85.1	76.0	$\mathbf{-0.13}^{\pm 0.4}$	$78.2^{\pm 3.3}$
+ weighted output	**86.1**	76.0	$-0.28^{\pm 0.2}$	$78.2^{\pm 0.6}$

UNIVERSITY
OF OSLO

Conclusion

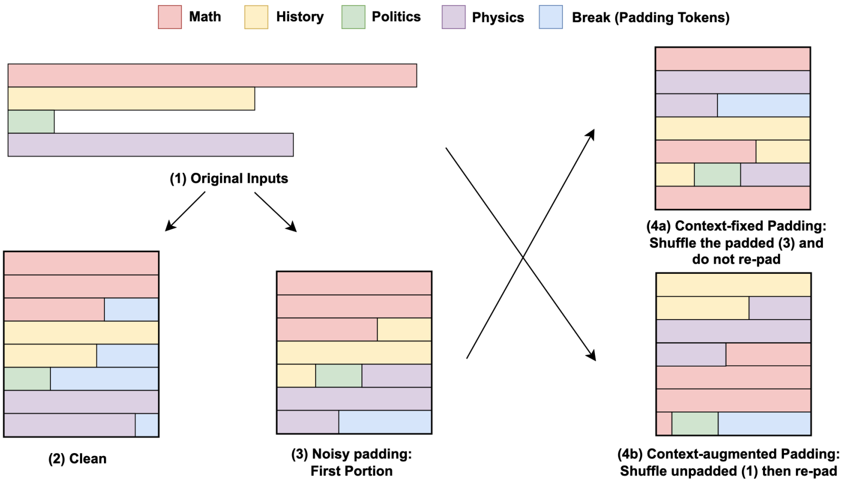
- Not all layers are equally as important.
- Focus on the previous layer for every layer and the embedding layer for the first five and last layers.
- Improved performance on (Super)GLUE and comparable on BLiMP.
- Potentially more linguistically biased.

Contents

Contextualizer

- **Paper**: Towards more Human-like Language Models based on Contextualizer Pretraining Strategy
- **Authors**: Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed
- Goals: Avoid the "contextualization trap", or always exposing the knowledge of a domain surrounded by the knowledge of that same domain.

Main Diagrams



Key Takeaways

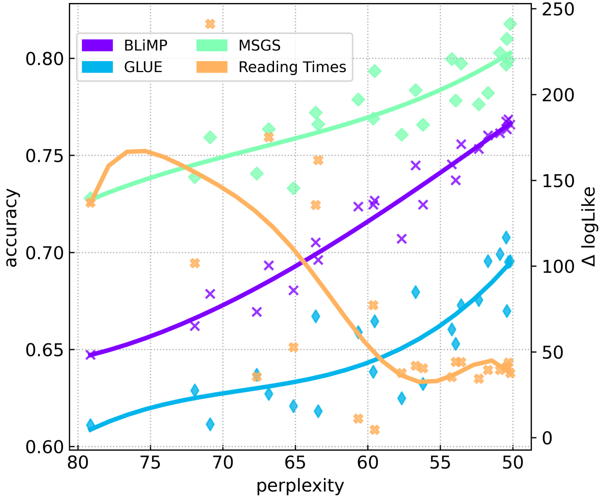
- First shuffling the data and then concatenating and padding it (4b) leads to substantial improvements.
- Doing a round of clean data before and after shows little gains.
- Works better for the 100M dataset than the 10M dataset.
- Potentially leads to models learning less shortcuts.
- BLiMP results on par with BERT and 1.2% lower than RoBERTa.

Contents

Large GPT-like Models are Bad Babies

- **Paper**: Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures
- **Authors**: Julius Steuer, Marius Mosbach, and Dietrich Klakow
- **Goals**: Access whether GPT-like models can acquire formal and functional linguistic competence as well as being "cognitively plausible".

Main Diagrams



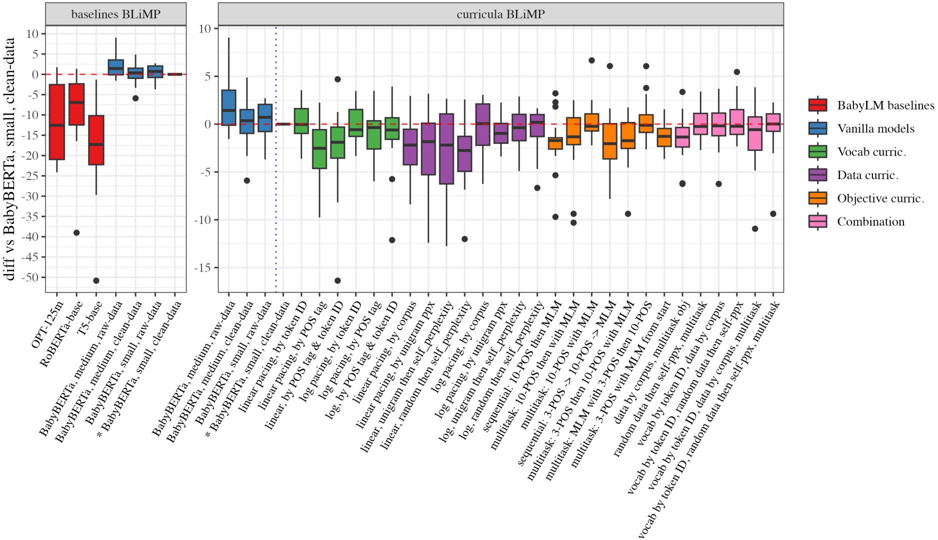
Key Takeaways

- GPT-like models can either acquire formal and functional linguistic competence or be "cognitively plausible" but not both.
- Best models on MSGS, GLUE and BLiMP are larger (>50M parameters).
- Best models for reading time are small (<5M parameters).
- Model size is not the only important factor for reading time, hidden size is also important.
- No or positive effect on reading time of training for multiple epochs.
- Using developmentally plausible datasets such as BabyLM is better for reading time.

CLIMB—Curriculum Learning for Infant-inspired Model Building

- **Paper**: CLIMB—Curriculum Learning for Infant-inspired Model Building
- **Authors**: Richard Diehl Martinez, Zébulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn
- **Goals**: Explore different types of curriculum learning to find one that improves LM performance.

Main Diagrams



UNIVERSITY
OF OSLO

Key Takeaways

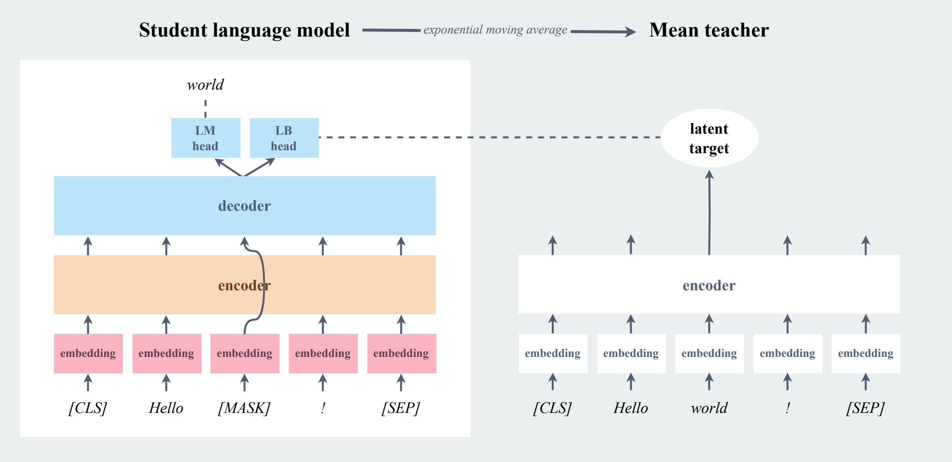
- Vocabulary curriculum: Different styles improve different tasks.
- Data curriculum: With multiple copora, ordering by difficulty can be useful.
- Objective curriculum: Multitask is better than sequentially changing objectives.
- Combining curricula: Shows potential on BLiMP, but not on other evaluation datasets.
- On small-corpora noisy data leads to better models than clean data.
- Overall, no curriculum method globally improves performance of the model, but can improve performance on specific tasks.

Contents

Mean BERTs make erratic language teachers

- **Paper**: Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings
- **Author**: David Samuel
- **Goals**: Test whether the success of latent supervision for computer vision can carry to NLP.

Main Diagrams



Key Takeaways

- Shows improvements on fine-tuning (Super)GLUE tasks.
- At the cost of performance on MSGS and mixed results on BLiMP.
- Latent supervision is great for computer vision, but results for NLP are more nuanced.
- Pre-training time is increased by 50%.

Lucas Georges Gabriel Charpentier

E-mail: lgcharpe@ifi.uio.no

BabyLM Challenge: What is it? Best models and Papers