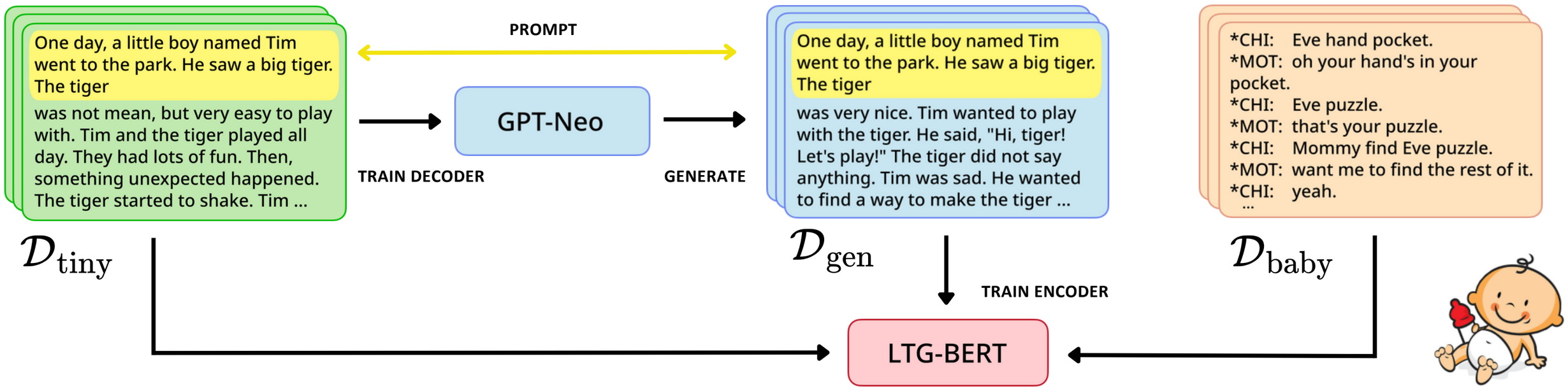
BERTtime Stories: Investigating the Role of Synthetic Story Data in Language Pre-training

Nikitas Theodoropoulos, Giorgos Filandrianos, Vasilis Lyberatos, Maria Lymperaiou, Giorgos Stamou
National Technical University of Athens,
nikitastheodorop@gmail.com, {marialymp, geofila, vaslyb}@ails.ece.ntua.gr, gstam@cs.ntua.gr

Main research questions

Q1: Can **data augmentation** with **synthetic story data** help pre-training?

Q2: Can LMs with **small training datasets** generate **high quality stories**?



Decoder Training

Use the **TinyStories** dataset (Dtiny) — a collection of short and simple stories, train GPT-Neo models for **Data Augmentation**. Train on subsets of 50M (Strict) and 5M (Strict-Small).

Data Generation

For each story in Dtiny truncate to 15%-30%, use GPT-Neo model to generate an **alternate completion**.

Encoder Training

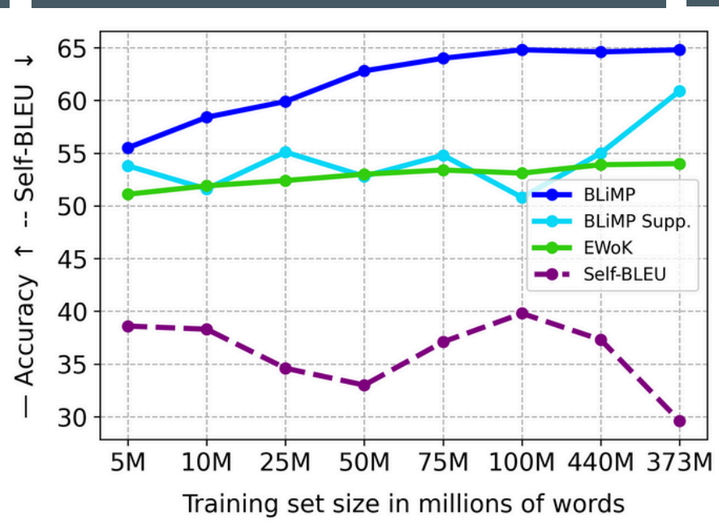
Combine: TinyStories (Dtiny), Generated Data (Dgen) and a subset of the BabyLM dataset (Dbaby) → **train LTG-BERT** model

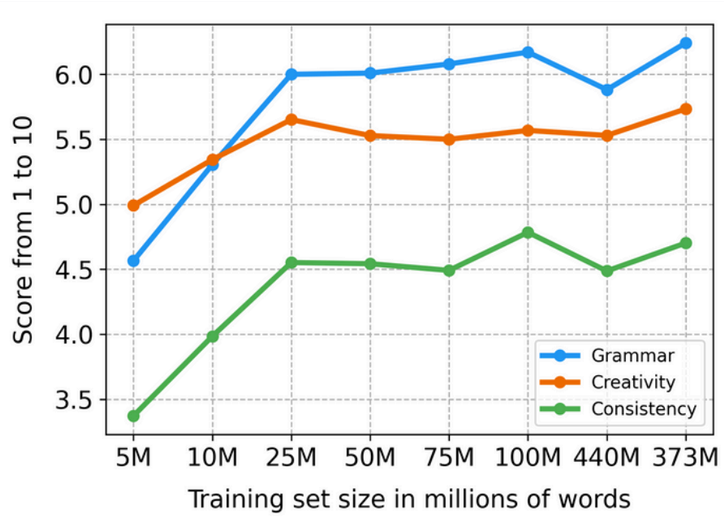
How much data is enough for good generations?

Train GPT-Neo models on various Dtiny sizes

- Evaluate BLiMP, BLiMP Supp., EWoK
- Evaluate Self-BLEU (diversity), LLM-eval: *Grammar, Creativity, Consistency* with plot

25M - 50M words are enough!





STRICT-SMALL

Model	Training Data	Total	BLiMP	Supp.	EWoK	GLUE	Avg.
LTG-BERT	baby-10M	10M	60.6	60.8	63.1	60.3	61.2
BabyLlama	baby-10M	10M	69.8	59.5	50.7	63.3	60.8
LTG-BERT (**ours**)	baby-10M	10M	62.8	**63.7**	66.2	71.0	**65.9**
	tiny-10M	10M	59.8	54.2	**67.0**	67.0	62.0
	+gen-greedy	20M	58.7	57.8	63.8	67.1	61.9
	baby-5M + tiny-5M	10M	62.6	60.7	66.6	**71.2**	65.3
	+gen-greedy	15M	62.1	60.2	65.5	70.6	64.6
	+gen-nucleus-1	15M	62.5	62.3	63.9	69.5	64.6
	+gen-nucleus-1 † ⋆	15M	**63.2**	59.3	65.5	71.1	64.8
	+gen-nucleus-5	33M	62.4	60.1	65.8	69.4	64.4
	+gen-nucleus-10	56M	61.0	58.4	65.3	69.5	63.6

STRICT

Model	Training Data	Total	BLiMP	Supp.	Ewok	GLUE	Avg
LTG-BERT	baby-100M	100M	69.2	66.5	65.7	68.4	67.5
BabyLlama	baby-100M	100M	73.1	60.6	52.1	69.0	63.7
LTG-BERT (**ours**)	baby-100M	100M	64.0	**67.6**	62.8	**74.0**	**67.1**
	tiny-100M	100M	61.2	63.2	63.1	70.6	64.5
	+gen-greedy	200M	61.1	59.6	63.8	69.1	63.4
	tiny-50M + baby-50M	100M	65.5	65.6	62.5	71.0	66.2
	+gen-greedy	150M	**66.6**	63.3	**65.0**	71.8	66.7
	+gen-nucleus-1⋆	150M	65.6	65.0	64.6	72.7	67.0
	+gen-nucleus-1†	150M	65.2	63.5	64.3	72.6	66.4
	+gen-nucleus-5	350M	65.4	64.4	61.2	69.8	65.2
	+gen-nucleus-10	600M	63.7	63.3	64.5	69.5	65.3

† = balanced training ⋆ = submitted model

- BabyLM 10M and 100M have best performance
- Adding more synthetic data hurts performance
- Nucleus and Greedy sampling resulted in some gains:
  - Best BLiMP score (Strict-Small)
  - Best BLiMP and EWoK scores (Strict)
- Synthetic story data offered some modest gains
- Overall, adding generated data hurts performance
- High generation quality underscores potential