LLM designs a study plan to train a smaller LLM

You are an LLM teaching a smaller model everything you know: Multi-task pretraining of language models with LLM-designed study plans

Wiktor Kamzela *wiktor.kamzela@student.put.edu.pl*
Mateusz Lango *lango@ufal.mff.cuni.cz*
Ondřej Dušek *odusek@ufal.mff.cuni.cz*

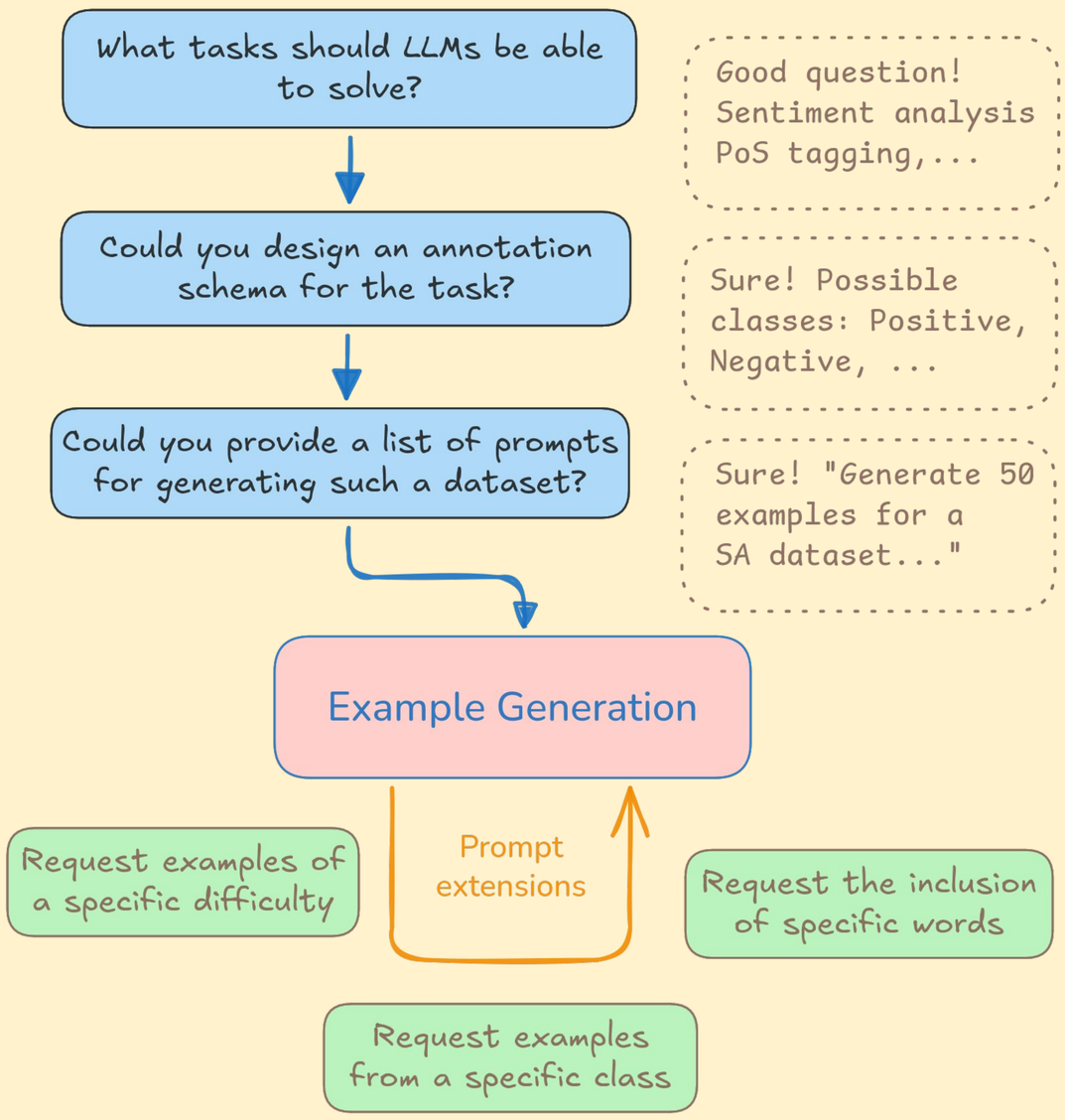
BabyLM Challenge

- **BabyLM Interaction track**
  - train an LLM using a **teacher LLM**
  - the teacher cannot reveal its weights, activation values,...
  - **limited generated data**

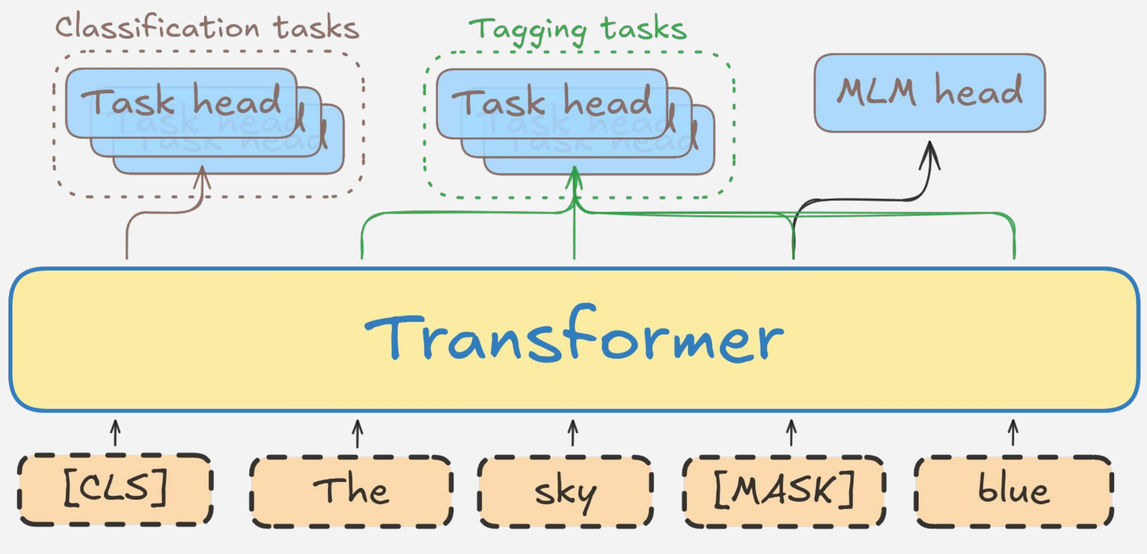
Our solution

- Model **trained *only* on synthetic data**
- **Teacher LLMs designs tasks** (classification, tagging, generation) to be learnt by student LLM
- Student is trained via multi-task pretraining (additional classification heads)

Synthetic Data Generation



Multi-task Pretraining



- The model optimizes a **loss function** which is a sum of:
  - standard MLM loss
  - all classificaffication tasks
  - all tagging tasks
- Labels for some tasks are not available for every instance
- Architecture: **149M ModernBERT** (39M also tested), context size 256
- Training by AdamW with 128 batch size

Methods

- **3 training datasets generated by Llama 3.1 8B**
  - Multi-task dataset
  - Text generation dataset
  - Vocabulary-controlled dataset
  - + standard Human-written BabyLM corpora
- **2 dataset sizes:** 1M and 10M words

Takeaways

- Training on small (1M) **synthetic data is more effective** than on 1M human data, both BLIMP and SuperGLUE
- On 10M data, **synthetic data enables higher SuperGLUE scores**

Results

D. size	Dataset	Epochs	BoolQ	MNLI	MultiRC	RTE	WSC	MRPC	QQP	BLiMP	S. GLUE	Average
1M	Text gen.	10	0.686	0.452	0.664	0.518	0.635	0.701	0.690	55.22	0.621	58.65
	Multi-task	10	0.713	0.444	0.669	0.554	0.615	0.730	0.715	56.80	0.634	60.11
	Vocab. c.	10	0.689	0.451	0.665	0.532	0.692	0.706	0.692	53.12	0.633	58.19
	Text gen.	100	0.689	0.459	0.666	0.554	0.654	0.706	0.713	55.56	**0.634**	59.50
	Multi-task	100	0.696	0.429	0.667	0.583	0.615	0.730	0.700	**57.82**	0.631	**60.48**
	Vocab. c.	100	0.691	0.448	0.665	0.547	0.635	0.721	0.704	56.09	0.630	59.55
	Human	10	0.691	0.442	0.660	0.597	0.635	0.721	0.699	54.38	**0.635**	58.94
		50	0.681	0.425	0.658	0.554	0.673	0.706	0.696	**57.66**	0.628	60.21
10M	Text gen.	10	0.708	0.494	0.665	0.525	0.654	0.745	0.744	61.84	0.648	63.32
	Multi-task	10	0.701	0.453	0.666	0.576	0.635	0.730	0.729	64.38	0.641	64.25
	Vocab. c.	10	0.704	0.485	0.674	0.568	0.635	0.745	0.740	61.78	0.650	63.39
	Text gen.	50	0.698	0.526	0.670	0.561	0.654	0.730	0.767	63.06	0.658	64.44
	Multi-task	50	0.707	0.445	0.673	0.547	0.615	0.696	0.733	**65.19**	0.631	64.14
	Vocab. c.	50	0.702	0.509	0.673	0.619	0.654	0.735	0.758	63.76	**0.664**	**65.10**
	Human	10	0.698	0.449	0.670	0.554	0.654	0.770	0.725	69.38	**0.646**	66.97
		50	0.694	0.458	0.668	0.576	0.654	0.730	0.745	**71.68**	**0.646**	**68.15**

Presented at BabyLM workshop @ EMNLP 2025, Suzhou, China

This research was co-funded by the European Union (ERC, NG-NLG,101039303).
This work used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

European Research Council
Established by the European Commission