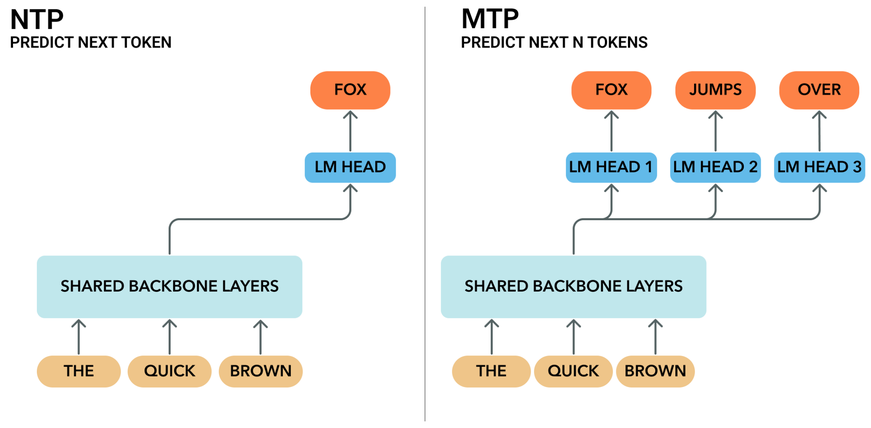
Babies Look Ahead: Multi-Token Prediction in Small LMs

Ansar Aynetdinov, Alan Akbik
Humboldt-Universität zu Berlin

Motivation



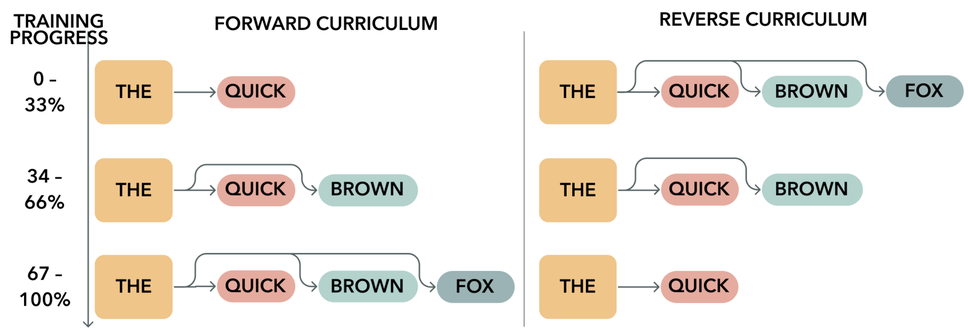
- **Multi-token prediction (MTP)** [1] has been proposed as an alternative to the **next-token prediction (NTP)** objective in training LLMs.
- Predicting multiple tokens at each prediction step, instead of only one, leads to **better downstream performance**.
- **Problem: small LMs (<7B)** are struggling with the MTP objective.

Our Idea: Leverage Curriculum Learning

Curriculum learning allows to **modulate the difficulty** of the MTP task by **controlling the number of predicted tokens** throughout the training.

We consider two types of curricula [2]:

- **Forward:** start with a vanilla NTP task, adding an additional token to the task every $\frac{n}{m}$ steps (**easy-to-hard**).
- **Reverse:** start with a full MTP task, dropping a token from the task every $\frac{n}{m}$ steps (**hard-to-easy**).



Experimental Setup

Model & MTP task: we consider a 130M model with a GPT-2 architecture trained on a **2-token** prediction task.

MTP module: $m-1$ auxiliary **LM heads** are dedicated to additional tokens.

Vocabulary size: we investigate the impact of a **large (16k) vs small (8k)** vocabulary on the downstream performance.

Results

Vocabulary Size	Objective	Curriculum	BLiMP (Acc.)	BLiMP Suppl. (Acc.)	EWoK (Acc.)	Entity Tracking (Acc.)	WUG Adj. Nom. (Acc.)	Eye Tracking (ΔR^2)	Self-paced Reading (ΔR^2)	Avg.
16k	NTP	-	**62.17**	**59.48**	49.79	13.74	59.50	10.59	4.13	37.06
	MTP	-	61.37	56.90	49.46	17.88	65.00	11.00	**4.30**	37.99
		Reverse	61.93	57.60	**50.22**	**18.60**	**66.00**	**11.17**	4.28	**38.54**
		Forward	61.51	58.29	49.73	13.40	60.00	10.15	4.00	36.73
8k	NTP	-	**61.91**	58.57	**49.51**	11.82	57.50	9.43	3.77	36.07
	MTP	-	61.25	56.81	49.12	16.25	**70.00**	8.92	3.61	**37.99**
		Reverse	61.36	56.61	49.22	**16.31**	66.50	8.58	3.58	37.45
		Forward	61.23	**59.10**	49.36	11.91	55.50	**9.48**	**3.83**	35.77

Table 3: Zero-shot evaluation of models using different tokenizer vocabulary sizes after training for 10 epochs on the 10M BabyLM dataset. **Best** scores are highlighted.

- **MTP objective acts as a regularizer** in the early epochs, preventing the SLMs from latching onto local patterns.
- **Switching to the NTP objective later on refines language representation**, while the opposite does not leads to additional benefits.

- **MTP improves downstream NTP performance** and forces SLMs to focus on patterns beyond local ones.
- **Larger vocabulary leads to more human-like text processing by SLMs** despite increased semantic complexity.

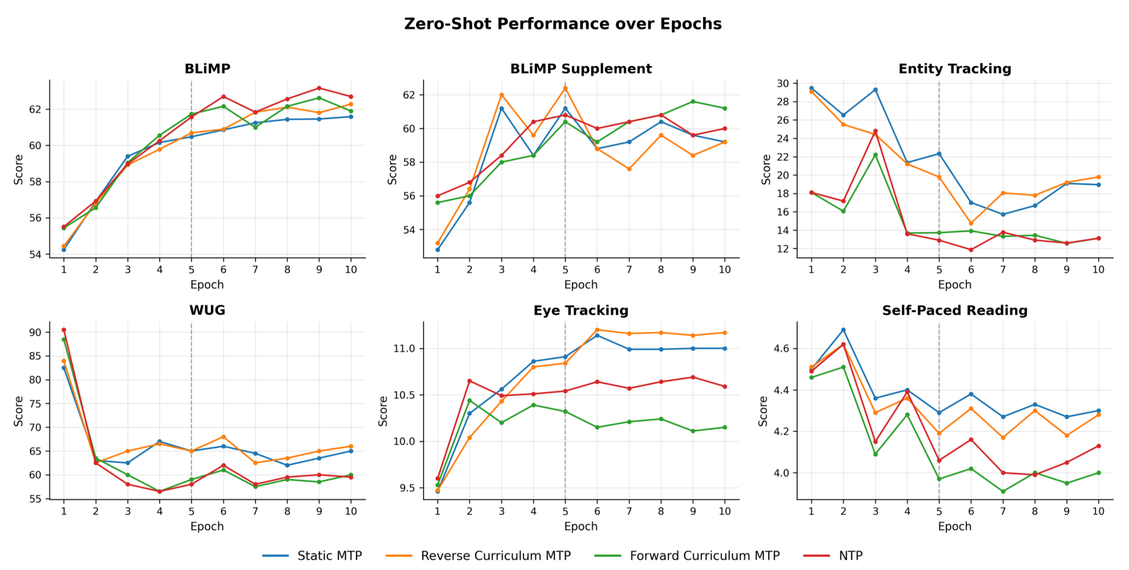


Figure 3: Zero-shot evaluation over epochs. The dotted line at epoch 5 indicates the switch in the training objective for models trained with either of the objective curricula. Tokenizer vocabulary size: **16K**.

References & Acknowledgments: [1] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. [2] Ansar Aynetdinov and Alan Akbik. 2025. Pre-Training Curriculum for Multi-Token Prediction in Language Models.
The authors are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Emmy Noether grant "Eidetic Representations of Natural Language" (project number 448414230). Further, Alan Akbik is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy "Science of Intelligence" (EXC 2002/1, project number 390523135).
Special thanks to Polina Aynetdinova for contributing the diagram designs for this poster.