# Pretraining language models with **artificial languages** and **LoRA adapters** enables **data-efficient** learning

## Pretraining Language Models with LoRA and Artificial Languages

**Nalin Kumar** *nkumar@ufal.mff.cuni.cz*
**Mateusz Lango** *lango@ufal.mff.cuni.cz*
**Ondřej Dušek** *odusek@ufal.mff.cuni.cz*

Charles University
ÚFAL

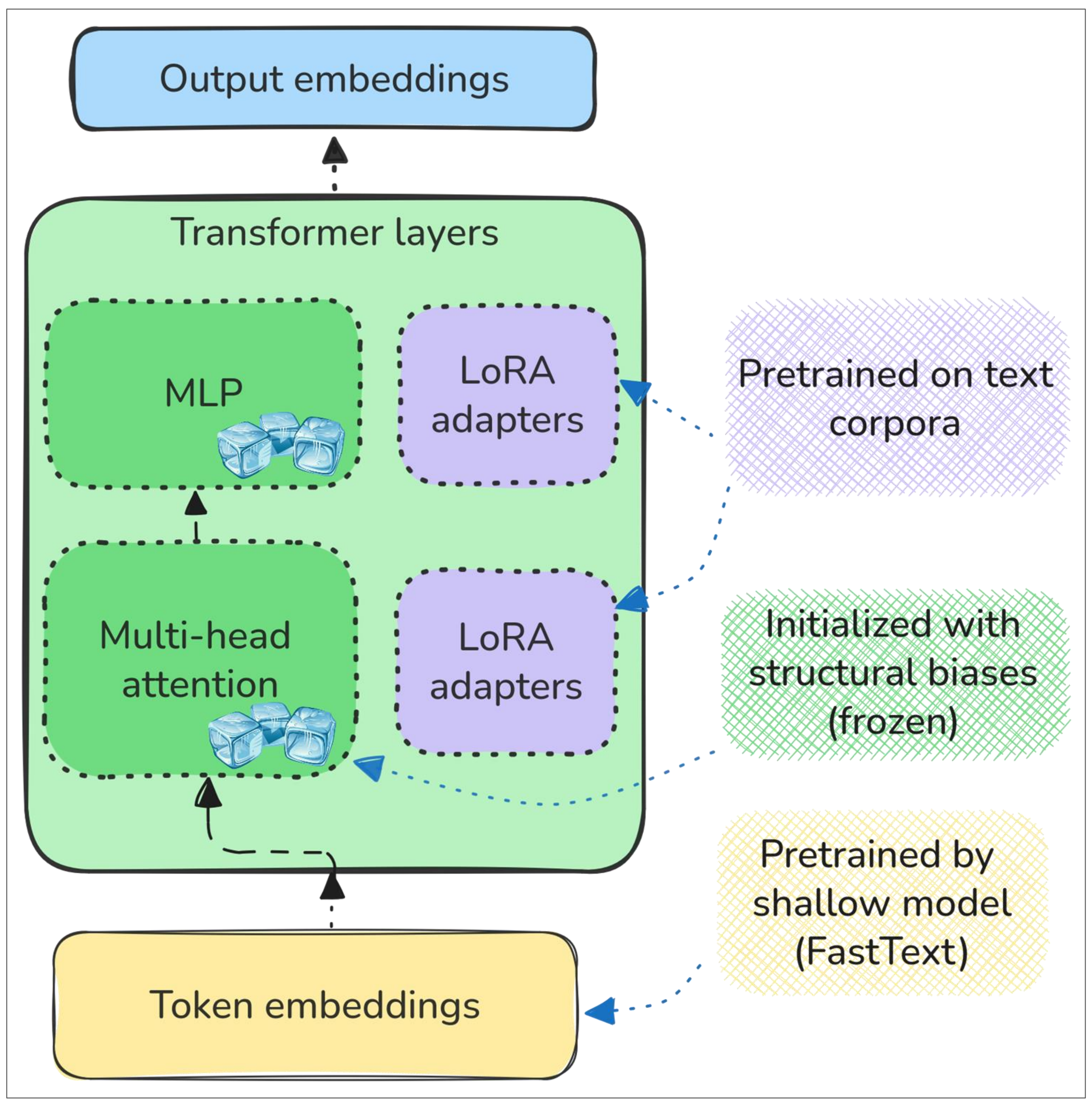
### Motivation

- LLMs require massive data unlike humans
- Current works focus on scaling up model and data size
- Need data-efficient methods — earlier works suggest non-linguistical data improves fine-tuning on NLP tasks

### Overview

- **Parameter-efficient pretraining** for language acquisition from limited data
- Combining **shallow embeddings**, **artificial languages**, and **LoRA adapters** for efficient pretraining

### Approach


Output embeddings
Transformer layers
MLP
LoRA adapters
Multi-head attention
LoRA adapters
Pretrained on text corpora
Initialized with structural biases (frozen)
Pretrained by shallow model (FastText)
Token embeddings

1. **Pretrained Embeddings**
   - Initialize token embeddings with FastText skip-gram model
   - Trained on same text corpus
   - Provides better surface-level lexical representation
2. **Artificial Language (AL) Pretraining**
   - Pretrain on NEST and CROSS artificial languages (Papadimitriou and Jurafsky, 2023)
   - Induces structural biases
3. **Parameter-Efficient Pretraining**
   - Train using LoRA low-rank adapters

### Experimental Setup

- **Dataset:** BabyLM 10M Corpus & AL — 20k samples (512 tokens)
- **Evaluation Metrics:** BLiMP, EWoK
- **Training Details**
  - FastText embed. (dim. 768) — using skip-gram for 5 epochs
  - Pretraining on AL for 25 epochs (equi. to 0.43 of 10M corpus)
  - LoRA training for 10 epochs

### Findings

- FastText initialization improves performance
- CROSS-language pretraining helps
- Increasing LoRA rank improves results

| Model | | | | | | |
|---|---|---|---|---|---|---|
| **Embed Init.** | **AL Init.** | **Pretraining** | **BLiMP** | **Supp.** | **EWoK** | **Avg** |
| Random | None | None | 54.91 | 47.25 | 50.09 | 50.75 |
| FastText | CROSS | None | 57.51 | 50.05 | **50.47** | 52.67 |
| FastText | NEST | None | 52.25 | 49.13 | 50.04 | 50.47 |
| Random | None | Standard | 56.26 | 48.48 | 50.09 | 51.61 |
| Random | None | LoRA (16) | 53.09 | 46.25 | 49.97 | 49.77 |
| Random | CROSS | LoRA (16) | 52.66 | 45.32 | 50.11 | 49.36 |
| FastText | CROSS | LoRA (16) | 58.18 | 51.98 | 50.38 | 53.51 |
| FastText | CROSS | LoRA (64) | 58.55 | 50.49 | 50.43 | 53.15 |
| FastText | CROSS | LoRA (128) | **60.96** | 51.27 | 50.25 | 54.16 |
| FastText | CROSS | LoRA (256) | 60.20 | **53.21** | 50.10 | **54.50** |

### Artificial Languages

#### (NEST)

- Vocabulary: opening and closing tokens ($p_{open}$ = 0.49, $p_{close}$= 0.51)
- Text generation: left to right
- If a closing token is selected, the most recently unmatched opening token is closed
- e.g. 1 ( 24 ( 24 ) 67 ( 39 ( 39 ) 67 ) 1 )

#### (CROSS)

- Similar vocabulary as NEST
- If a closing token is selected, any unmatched token can be closed
- e.g. 1 ( 24 ( 67 ( **24** ) 39 ( 39 ) 1 ) **67** )

***Presented at BabyLM Challenge 2025, EMNLP***

*This work was supported by the European Research Council (Grant agreement No. 101039303, NG-NLG) and Grant Agency of Charles University (Grant No. 302425), and used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).*

**HF Repo: *knalin55/CUNI_FT-CROSS-LoRA***

European Research Council
Established by the European Commission