What Causes Knowledge Loss in Multilingual Language Models?

Maria Khelli[1]—khelli07.id@gmail.com
Samuel Cahyawijaya[2]—samuelcahyawijaya@cohere.com
Ayu Purwarianti[1]—ayu@informatika.org
Genta Indra Winata[3]—genta.winata@capitalone.com

[1] Institut Teknologi Bandung
[2] Cohere
[3] Capital One

TL;DR: Parameter sharing helps mitigate forgetting, especially for Latin-script languages, while non-Latin scripts remain more vulnerable.

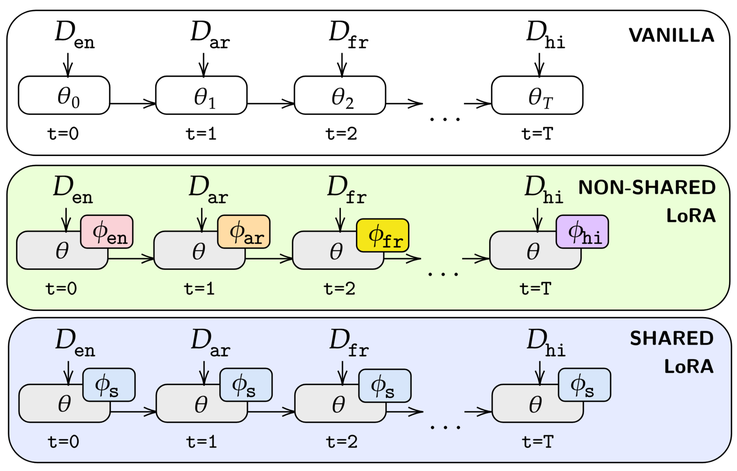
Introduction

In real-world scenarios, data often arrives sequentially rather than all at once, requiring models to learn continuously (Rolnick et al., 2019).

Sequential training can lead to catastrophic forgetting—a degradation in performance as models learn new tasks, especially in multilingual settings.

We investigate how parameter sharing and linguistic differences across 52 languages impact catastrophic forgetting, analyzing representational shifts and knowledge erosion beyond model parameters.

Proposed Method



Our pipeline for various approaches in lifelong learning. We also add **Multi** as or upper bound. Mechanism: a single model (or adapter) is **optimized on all languages simultaneously**, over the entire dataset *D*.

A single model is trained incrementally, with all parameters updated for each new language.

Each language gets its own separate LoRA adapter, while the main model (θ) is frozen.

A single, shared LoRA adapter is trained, while the main model's parameters (θ) are frozen.

Setup

MASSIVE (FitzGerald et al., 2023)
Task: Slot-filling

XLM-R (Conneau et al., 2020) and **E5** (Wang et al., 2024)

Metrics

Cross-lingual Transfer

$$CFT = \frac{1}{T-1}\sum_{i=1}^{T-1}\bar{X}_i \quad \text{where} \quad \bar{X}_i = \frac{1}{T-i}\sum_{j=i+1}^{T} R_{i,j}$$

$$CBT = \frac{1}{T-1}\sum_{i=1}^{T-1}(R_{T-1,i} - R_{i,i})$$

where
R_{ij} = F_1 on task t_j after training on the last sample from task t_i
T = total tasks (languages) used = 52

Multi-Hop Transfer

$$MFT_h = \frac{1}{|L|}\sum_{l\in L}(\mathcal{P}_{i+h} - \mathcal{P}_{i-1})$$

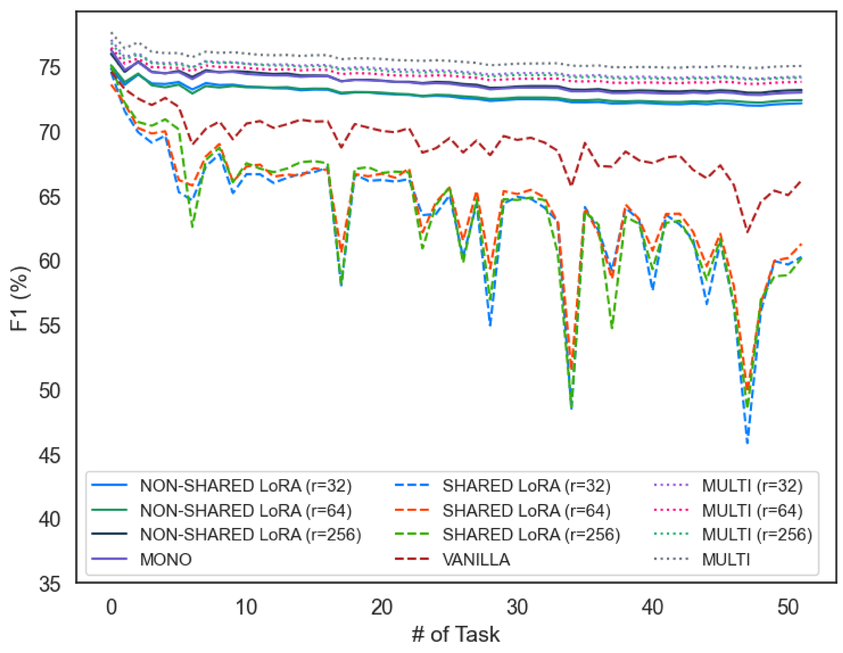
$$MBT_h = \frac{1}{|L|}\sum_{l\in L}(\mathcal{P}_{i} - \mathcal{P}_{i-h-1})$$

where
$\mathcal{P}_i$ = average F_1 on tasks seen up to step i
L = total number of languages (52)
h = number of hops

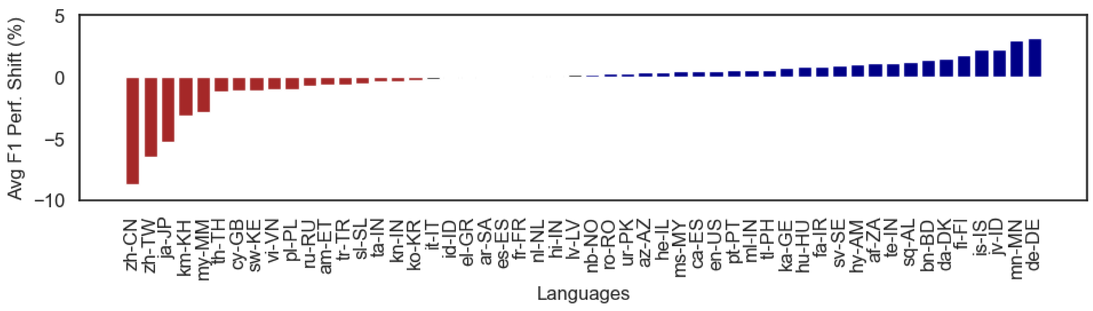
Results

Method	Params (M)	F1 (%)	Language Vitality		
			Low	Mid	High
MULTI	278.04	**75.03**	75.42	**75.84**	72.63
$r = 32$	5.36	74.19	74.27	**75.17**	71.83
$r = 64$	10.72	73.79	74.00	**74.56**	71.73
$r = 256$	42.86	74.11	74.16	**74.83**	72.41
MONO	14,458.27	72.98	73.66	**74.11**	69.43
VANILLA	278.04	66.16	65.70	**67.65**	63.46
SHARED LoRA					
$r = 32$	5.36	60.24	59.35	**62.34**	56.75
$r = 64$	10.72	**61.26**	60.55	**63.37**	57.48
$r = 256$	42.86	60.16	59.06	**62.15**	57.22
NON-SHARED LoRA					
$r = 32$	278.04	72.14	72.42	**73.39**	68.89
$r = 64$	557.19	72.38	72.55	**73.48**	69.65
$r = 256$	2,228.75	**73.16**	73.82	**74.26**	69.73

Comparison of methods based on trainable parameters and averaged F1 (%). Lower trainable the better.



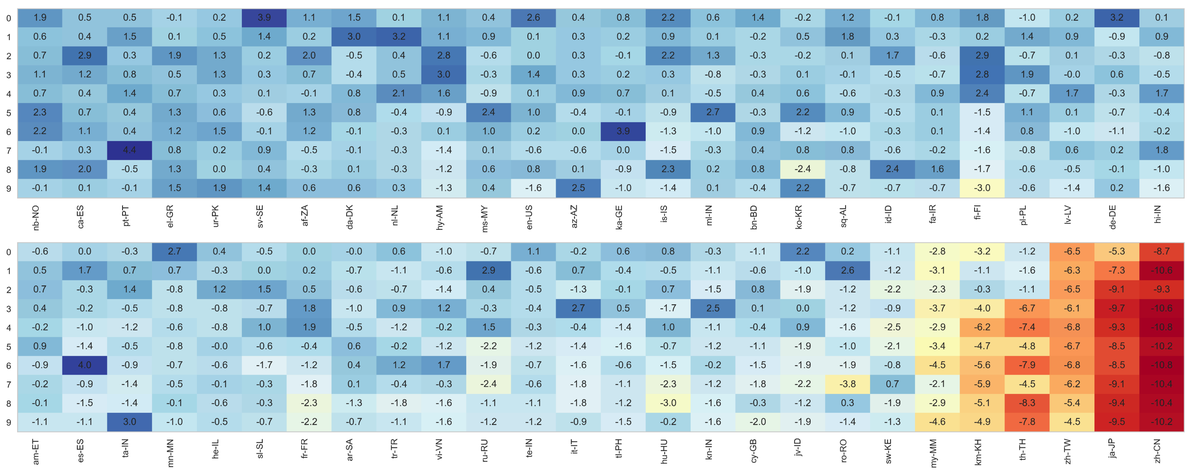
F1 average results after training each language over the time.



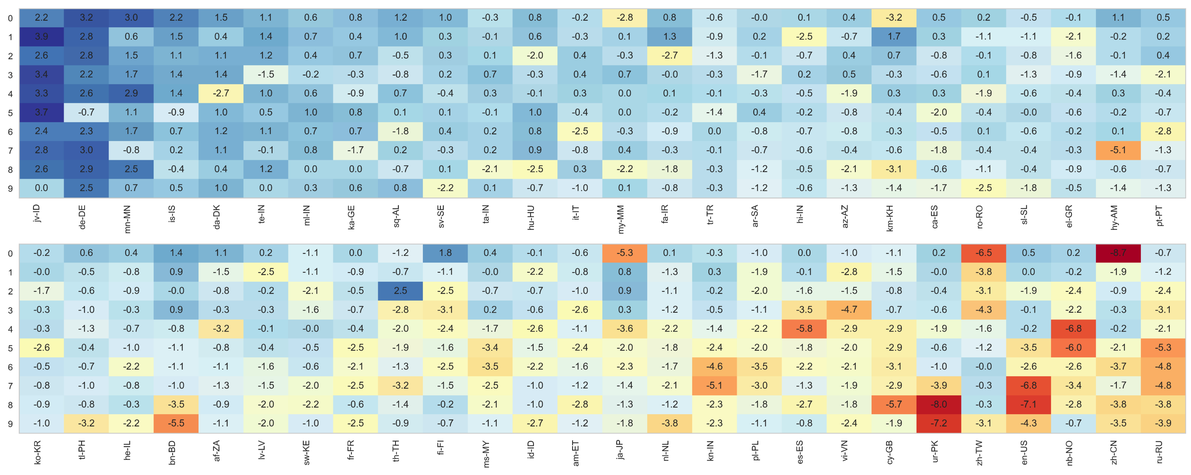
Average F1 shift (%): Chinese (zh-CN) exhibits the most significant perf. decline, while German (de-DE) serves as the most effective donor language.

Method	CBT	CFT
VANILLA	**-0.08**	**0.55**
SHARED LoRA		
$r = 32$	-0.13	0.52
$r = 64$	-0.12	0.53
$r = 256$	-0.14	0.51

CBT and CFT metrics for VANILLA and SHARED LoRA models. Higher values indicate better performance.



Multi-hop Backward Transfer (MBT), illustrates how training on later languages affects earlier ones over increasing hop distances (y-axis: 0–9). Orders of the language is sorted descending (read from top-left to bottom-right) based on its average over all hops.



Multi-hop Forward Transfer (MFT), represents each language's ability to donate knowledge to subsequent tasks over increasing hop distances (y-axis: 0–9). Orders of the language is sorted descending (read from top-left to bottom-right) based on its average over all hops.

Key Takeaways

Language impacts forgetting, script & tokenization matter: Non-latin scripts (Chinese, Japanese, Thai, Khmer) tend to cause more forgetting; Latin-script languages transfer better.

Parameter sharing trade-offs: per-language LoRA adapters offer strong performance with efficiency.

Transfer depends on order: Training sequence affects cross-lingual transfer; some languages help or hurt based on position.