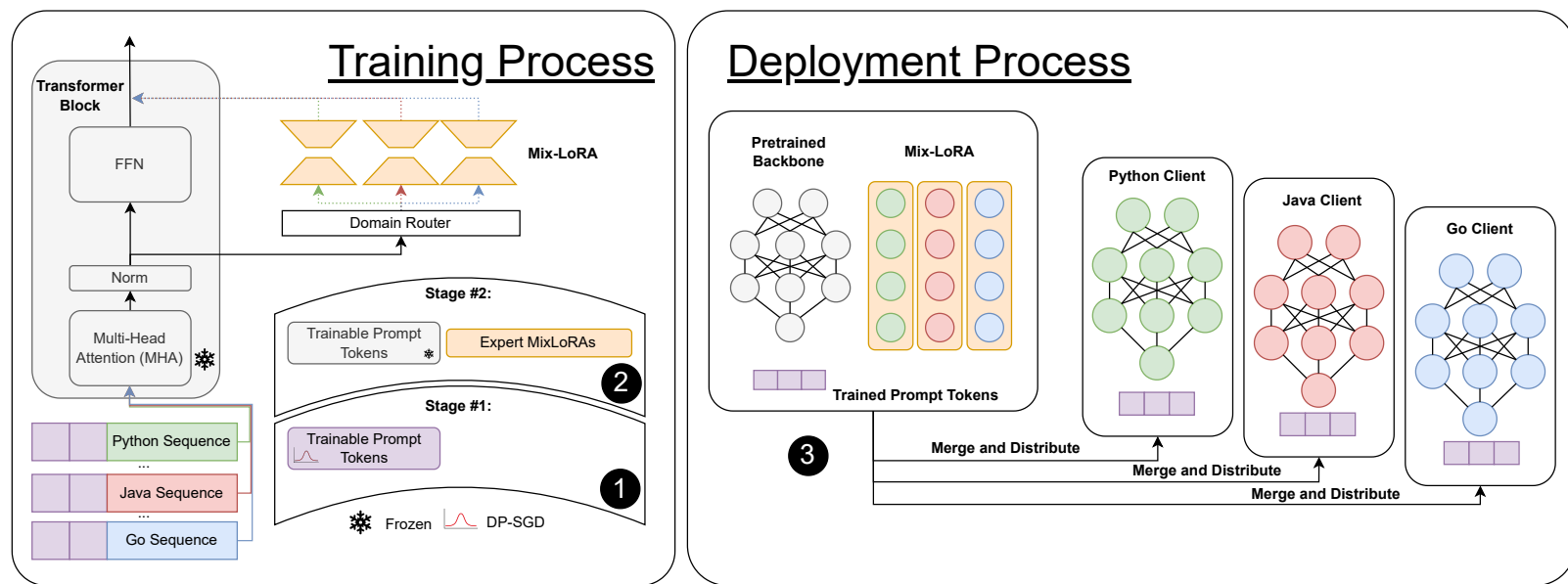


# NoEsis: Differentially Private Knowledge Transfer in Modular LLM Adaptation

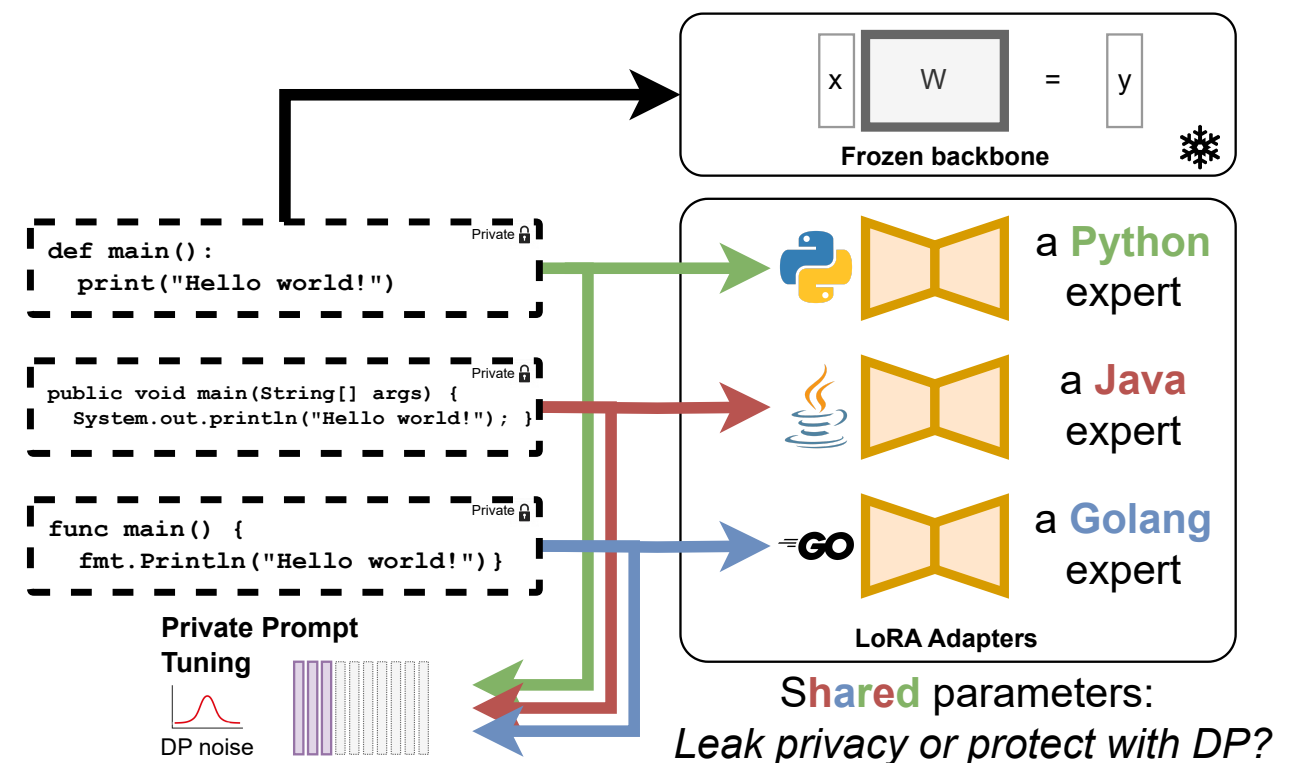


Rob Romijnders<sup>1</sup>, Stefanos Laskaridis<sup>\*2</sup>, Ali Shahin Shamsabadi<sup>2</sup>, Hamed Haddadi<sup>2</sup>

<sup>1</sup>Research done as internship  
<sup>2</sup>Brave Research



LLMs at scale are running out of training data. Tapping into private data silos is the promising next step, but privacy concerns are paramount [2,3]. NoEsis tackles this setting by adopting a hybrid PEFT paradigm to achieve the properties of **modularity**, **privacy**, and **knowledge transfer**.



Overview of modularity: in each domain, privacy w.r.t. local data is required, but knowledge transfer between domains is desired via shared parameters.

Model	$\epsilon$	Modular	Private	Transfer	PEFT	Python	Java	Go
(i) Share Nothing	0.0	✓	✓	✗	✓	68.31	60.19	64.17
(ii) Solo (separate models)	1.0	✗	✓	✗	✗	30.19	14.84	4.84
(iii) Monolithic Fine-tuning (Abadi et al., 2016)	1.0	✗	✓	✓	✗	36.05	23.54	18.34
(iv) Single Common Adapter (rc=512)* (Yu et al., 2021b)	1.0	✗	✓	✓	✓	48.21	36.16	27.24
(v) Prompt-Tuning Only (pt32) <sup>†</sup> (Duan et al., 2023)	1.0	✗	✓	✓	✓	35.03	24.29	9.67
(vi) NoEsis (pt32) <sup>†</sup>	1.0	✓	✓	✓	✓	<b>69.14</b>	<b>61.18</b>	<b>66.53</b>

\*rc, rank  $r$  of the common LoRA; <sup>†</sup>pt: number of trainable prompt tokens

Experimental comparison: NoEsis uniquely achieves high accuracy while obtaining DP at  $\epsilon = 1.0$  and enabling knowledge transfer across domains.

(Knowledge transfer is the accuracy increase over "Share Nothing," which does not have shared parameters between domains.)

$$\text{NoE} = \text{DP-SGD}(\text{Prompt-Tuning}(X_1, \dots, X_K)) + \text{SGD}(\text{Mix}\{\text{LoRA}(X_1), \dots, \text{LoRA}(X_K)\})$$

DP: A randomized algorithm  $A(\cdot)$  is  $(\epsilon, \delta)$ -differentially private if the following holds for any two adjacent data sets  $D, D'$ , and for any subset  $\Phi$  of outputs:

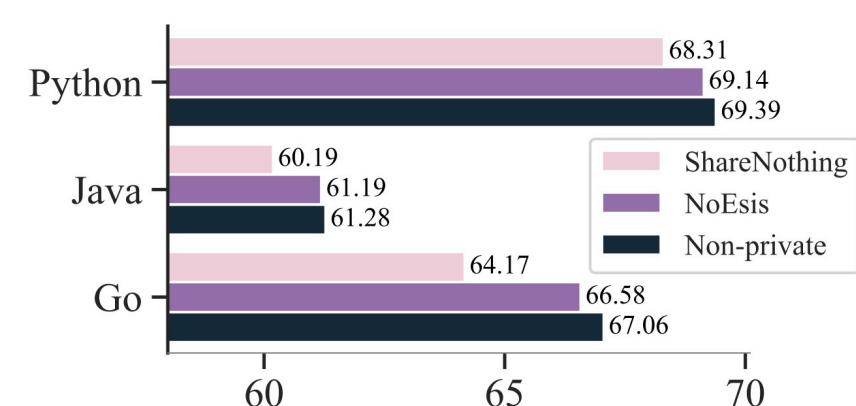
$$\Pr[A(D) \in \Phi] \leq e^\epsilon \Pr[A(D') \in \Phi] + \delta$$

where  $D = \{d_i\}_{i=1}^N$  is a dataset that contains  $N$  documents. Two datasets  $D$  and  $D'$  are adjacent when they are identical except for one document in any domain being removed.

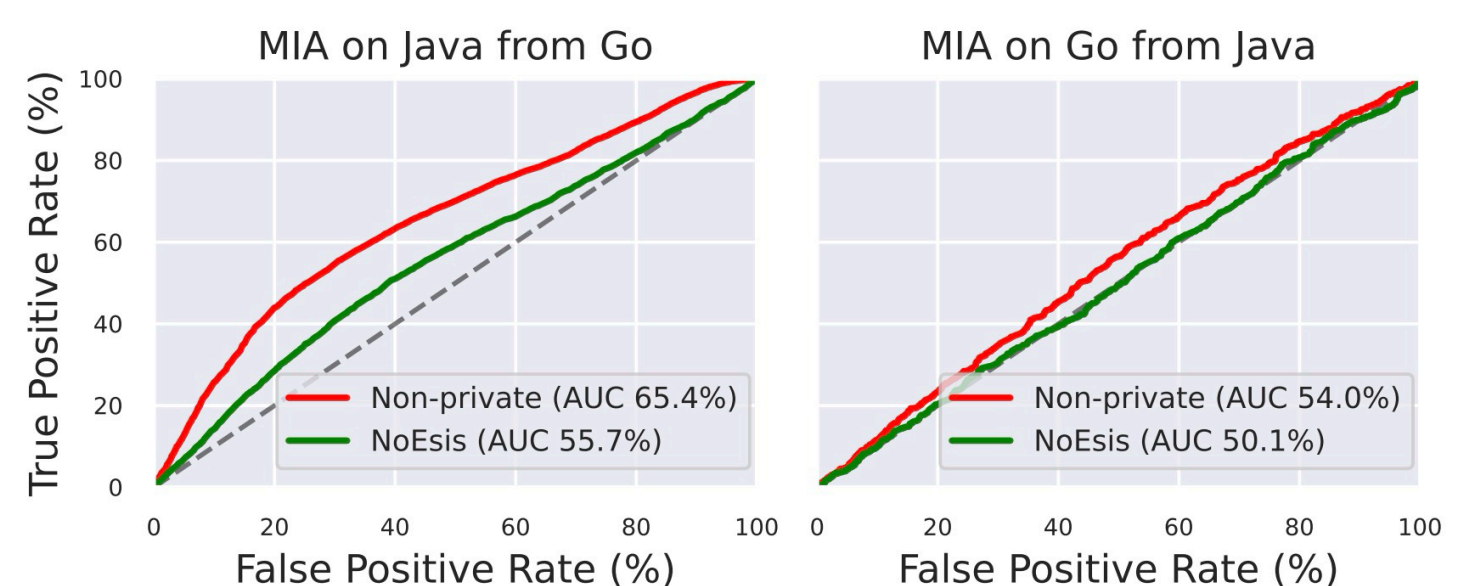
Let  $W \in \mathbb{R}^{p \times q}$  be the weight matrix of a linear layer in the FFN of a transformer with parameters  $\theta$ . We decompose this matrix:

$$W = W + \alpha \sum_{k=1}^K B^{(k)} A^{(k)}$$

where  $A^{(k)} \in \mathbb{R}^{r \times q}$  and  $B^{(k)} \in \mathbb{R}^{p \times r}$  are the trainable low-rank matrices for the  $k$ -th domain-specific adapter, out of  $K$  domains.



Between the results of a nonshared model, which is the baseline, and a non-private model, NoEsis has knowledge transfer and, under privacy constraints, bridges the accuracy gap by more than 77%.



We run a novel **cross-domain** Membership Inference Attack on the shared parameters of a Mixture-of-LoRA model. NoEsis maintains knowledge transfer and reduces vulnerability to this attack, for example, from 65.4 AUC to 55.7 AUC.