Disentangling Latent Shifts of In-Context Learning with Weak Supervision

Josip Jukić Jan Šnajder

TakeLab, University of Zagreb

NeurIPS 2025

Motivation

- **In-Context Learning (ICL) is unstable**: performance varies with demonstration choice/order.
- **ICL is inefficient**: long prompts increase latency and cost; limited by context window.
- **Idea**: Treat ICL as weak supervision. Learn the *latent shift* induced by demos and **store it in an adapter**.

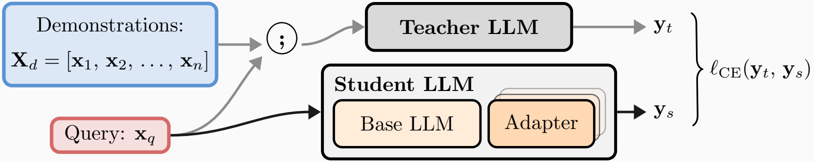
Linearized Attention View

$$\mathbf{f}_{\mathrm{AH}}(\mathbf{x}_q^{(t)}) \approx \underbrace{\mathbf{W}_{\mathrm{ZS}}\,\mathbf{q}^{(t)}}_{\text{zero-shot}} + \underbrace{\Delta\mathbf{W}_{\mathrm{ICL}}\,\mathbf{q}^{(t)}}_{\text{latent shift from demos}}$$

- Prior analyses often assume **linear attention**, neglecting nonlinear/residual dynamics.
- **Goal**: learn $\Delta\mathbf{W}_{\mathrm{ICL}}$ into a compact adapter.

$\mathbf{x}_q^{(t)}$ – query token at step t
$\mathbf{q}^{(t)}$ – query vector
$\mathbf{W}_{\mathrm{ZS}}$ – zero-shot weights
$\Delta\mathbf{W}_{\mathrm{ICL}}$ – demo-induced shift.

Weak Supervision



$$\mathcal{L} = \sum_{\mathbf{x}_q \in \mathcal{D}_{\text{unlab}}} \ell_{\text{CE}}\Big(\mathbf{f}_{\text{teacher}}\big([\mathbf{X}_d; \mathbf{x}_q]\big), \mathbf{f}_{\text{student}}\big(\mathbf{x}_q\big)\Big)$$

- Teacher conditions on **demos + query**; student uses **query only**.
- Student adapter learns to **match teacher logits** $\Rightarrow$ weakly supervised.

Generalization Experiments

Method	GLUE						MMLU
	RTE	QNLI	MNLI	COLA	MRPC	QQP	MISC
n-shot	$75.1_{6.5}$	$77.0_{5.5}$	$68.0_{3.0}$	$58.5_{4.0}$	$74.0_{2.5}$	$70.0_{3.0}$	$84.0_{4.0}$
PBFT	$73.2_{3.8}$	$77.8_{6.0}$	$67.4_{3.5}$	$56.5_{3.0}$	$72.0_{2.0}$	$68.0_{2.5}$	$83.5_{4.5}$
Batch-ICL	$77.8_{4.7}$	$78.0_{6.0}$	$70.9_{3.5}$	$59.8_{3.7}$	$75.2_{2.2}$	$72.5_{2.7}$	$81.0_{2.5}$
WILDA	$\mathbf{86.0}_{0.6}$	$\mathbf{81.4}_{2.2}$	$\mathbf{73.1}_{2.0}$	$\mathbf{64.3}_{2.2}$	$\mathbf{77.7}_{1.5}$	$\mathbf{73.1}_{1.8}$	$\mathbf{88.0}_{2.2}$

ID generalization accuracy for **Llama 3 (8B)** in the 16-shot setup.

- WILDA **achieves the strongest ID generalization**, outperforming standard ICL and related methods.
- **Highly stable**: variance is reduced compared to standard ICL.
- **Strong OOD generalization**: WILDA maintains high accuracy and low variance when evaluated on near-OOD GLUE pairs.

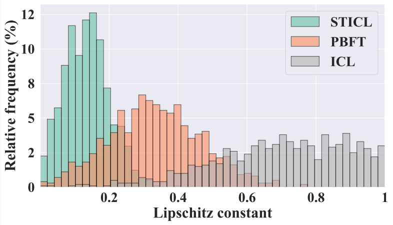
KNOWLEDGE FUSION VIA ADAPTER ARITHMETIC

		GLUE						MMLU
Demos	**Method**	RTE	QNLI	MNLI	COLA	MRPC	QQP	MISC
32	*n*-shot	$75.3_{3.2}$	$77.7_{2.9}$	$69.1_{1.9}$	$58.3_{1.5}$	$76.4_{2.2}$	$74.2_{1.9}$	$84.5_{2.1}$
32	WILDA	$\mathbf{87.9}_{0.6}$	$\mathbf{83.1}_{0.9}$	$74.0_{1.1}$	$64.6_{1.2}$	$\mathbf{79.4}_{0.6}$	$\mathbf{74.8}_{1.5}$	$89.0_{0.4}$
2×16	WILDA	$87.1_{1.6}$	$81.5_{5.0}$	$\mathbf{75.5}_{2.5}$	$\mathbf{68.4}_{1.8}$	$78.5_{1.4}$	$74.1_{1.6}$	$\mathbf{89.5}_{2.0}$

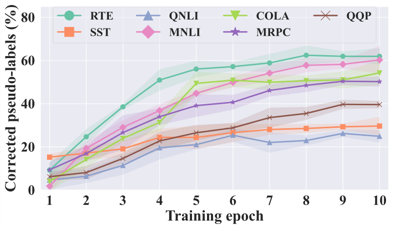
ID generalization accuracy for **Llama 3 (8B)** with fused demonstrations.

- **Adapter arithmetic** merges latent shifts from multiple subsets without retraining.
- **Stable fusion:** variance remains low as subsets increase.
- Enables **scalable task composition** – beyond context-window limits.

Weak-to-Strong Generalization I



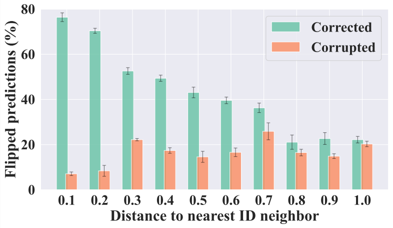
(a) Approximated Lipschitz constants



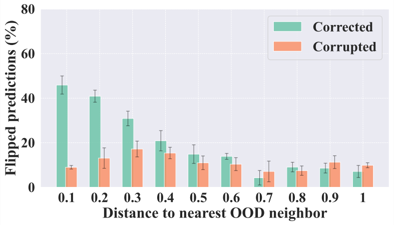
(b) Pseudo-label correction over epochs

- WILDA achieves strong local stability
- **Pseudo-label** corrections steadily increase across epochs

Weak-to-Strong Generalization II



(c) ID corrected/corrupted rates



(d) OOD corrected/corrupted rates

- Correction rates fall sharply with distance to the nearest correctly pseudo-labeled neighbor $\rightarrow$ **coverage expansion**.
- The same pattern appears on OOD data $\rightarrow$ WILDA generalizes corrections beyond the ID manifold.