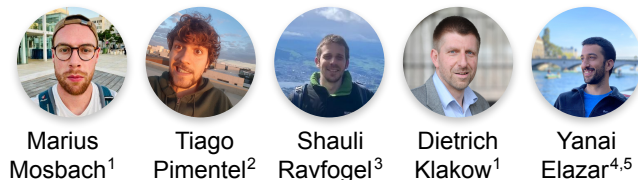


# Few-shot Fine-tuning vs. In-context Learning

## A Fair Comparison and Evaluation



Marius Mosbach<sup>1</sup>

Tiago Pimentel<sup>2</sup>

Shauli Ravfogel<sup>3</sup>

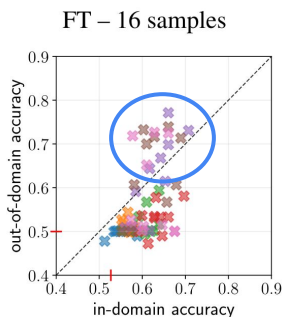
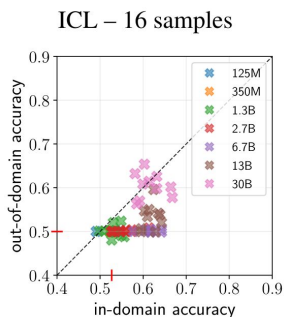
Dietrich Klakow<sup>1</sup>

Yanai Elazar<sup>4,5</sup>

1 Test generalization for models adapted via **fine-tuning (FT)** and **in-context learning (ICL)**

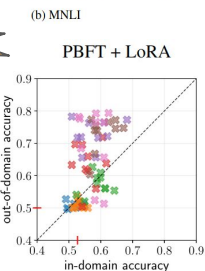
2 Findings **generalize** to other models (Pythia) and FT via LoRA

3 **High-level comparison:**



		FT							FT							
		125M	350M	1.3B	2.7B	6.7B	13B	30B	125M	350M	1.3B	2.7B	6.7B	13B	30B	
ICL	125M	-0.00	0.01	0.02	0.03	0.12	0.14	0.09	125M	-0.00	0.00	-0.02	0.01	0.10	0.11	0.07
	350M	-0.00	0.01	0.02	0.03	0.12	0.14	0.09	350M	-0.00	0.00	0.02	0.00	0.10	0.11	0.07
	1.3B	-0.00	0.01	0.02	0.03	0.12	0.14	0.09	1.3B	-0.01	-0.00	0.01	0.01	0.09	0.11	0.07
	2.7B	-0.00	0.01	0.02	0.03	0.12	0.14	0.09	2.7B	-0.01	-0.00	0.01	0.01	0.09	0.10	0.06
	6.7B	-0.00	0.01	0.02	0.03	0.12	0.14	0.09	6.7B	-0.01	-0.01	0.01	0.00	0.09	0.10	0.06
	13B	-0.04	-0.02	-0.01	-0.00	0.09	0.11	0.05	13B	-0.03	-0.03	-0.02	-0.00	0.07	0.08	0.04
	30B	-0.11	-0.09	-0.08	-0.08	0.02	0.03	-0.02	30B	-0.07	-0.07	-0.05	-0.06	0.03	0.04	0.00

(a) RTE **Unfair!**



		FT				
		410M	1.4B	2.8B	6.9B	12B
ICL	410M	0.02	0.06	0.05	0.09	0.11
	1.4B	0.01	0.05	0.04	0.08	0.10
	2.8B	-0.03	0.01	-0.00	0.04	0.06
	6.9B	0.01	0.05	0.04	0.08	0.10
	12B	-0.03	0.01	-0.00	0.04	0.06

Significance tests: ICL > FT, FT > ICL

	FT	ICL
<b>Users</b>	Experts	Experts & Non-experts
<b>Interaction</b>	Pre-defined	Textual
<b>Reusability</b>	Medium	High
<b>Applicability to low-resource languages</b>	High	Limited
<b>Requires training</b>	Yes	No
<b>Number of samples</b>	Unlimited	0 to ?
<b>Variance</b>	High	High
<b>Well understood</b>	No	No

Fine-tuned models can generalize well out-of-domain

Approaches should be compared using models of the same size

Robust task adaptation remains a challenge!



uds-lsv/llmft