ID	title	topic	link
1	Exploring Transformers as Compact, Data-efficient Language Models	data efficiency	https://aclanthology.org/2023.conll-1.35/
2	InsCL: A Data-efficient Continual Learning Paradigm for Fine-tuning Large Language Models with Instructions	data efficiency	https://aclanthology.org/2024.naacl-long.37/
3	Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling	data efficiency	https://aclanthology.org/2024.acl-long.757/
4	How to Train Data-Efficient LLMs	data efficiency	https://arxiv.org/abs/2402.09668
5	STAR: Constraint LoRA with Dynamic Active Learning for Data-Efficient Fine-Tuning of Large Language Models	data & model efficiency	https://aclanthology.org/2024.findings-acl.209/
6	AutoPEFT: Automatic Configuration Search for Parameter-Efficient Fine-Tuning	model efficiency (longer paper, 13 pages!)	https://aclanthology.org/2024.tacl-1.29/
7	Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding	model efficiency (inference)	https://aclanthology.org/2024.findings-acl.456/
8	LRQuant: Learnable and Robust Post-Training Quantization for Large Language Models	model efficiency (inference)	https://aclanthology.org/2024.acl-long.122/
9	LoRA-drop: Efficient LoRA Parameter Pruning based on Output Evaluation	model efficiency (training)	https://aclanthology.org/2025.coling-main.371/
10	MoSLD: An Extremely Parameter-Efficient Mixture-of-Shared LoRAs for Multi-Task Learning	model efficiency (training)	https://aclanthology.org/2025.coling-main.111/
11	A Study of Parameter Efficient Fine-tuning by Learning to Efficiently Fine-Tune	model efficiency (fine-tuning)	https://aclanthology.org/2024.findings-emnlp.929.pdf
12	RECOST: External Knowledge Guided Data-efficient Instruction Tuning	data efficiency (fine-tuning on synthetic data)	https://aclanthology.org/2024.findings-acl.648.pdf
13	Self-play fine-tuning converts weak language models to strong language models	data efficiency (fine-tuning on synthetic data)	https://arxiv.org/pdf/2401.01335
14	MEDUSA: Simple LLM inference acceleration framework with multiple decoding heads	model efficiency (inference)	https://arxiv.org/pdf/2401.10774
15	OneBit: Towards Extremely Low-bit Large Language Models	model efficiency (compression)	https://arxiv.org/pdf/2402.11295

16	Quantized Side Tuning: Fast and Memory-Efficient Tuning of Quantized Large Language Models	model efficiency (compression, PEFT)	https://aclanthology.org/2024.acl-long.1.pdf
17	FreeAL: Towards Human-Free Active Learning in the Era of Large Language Models	data efficiency	https://aclanthology.org/2023.emnlp-main.896
18	MELM: Data Augmentation with Masked Entity Language Modeling for Low-Resource NER	data efficiency	https://aclanthology.org/2022.acl-long.160/
19	Does the Order of Training Samples Matter? Improving Neural Data-to-Text Generation with Curriculum Learning	data efficiency	https://aclanthology.org/2021.eacl-main.61.pdf
20	Applying Natural Annotation and Curriculum Learning to Named Entity Recognition for Under-Resourced Languages	data efficiency	https://aclanthology.org/2022.coling-1.394.pdf
21	Do We Need to Create Big Datasets to Learn a Task?	data efficiency	https://aclanthology.org/2020.sustainlp-1.23/
22	NLP From Scratch Without Large-Scale Pretraining: A Simple and Efficient Framework	data efficiency	https://proceedings.mlr.press/v162/yao22c/yao22c.pdf
23	Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes	data & model efficiency	https://arxiv.org/pdf/2305.02301.pdf
24	ATTEMPT: Parameter-Efficient Multi-task Tuning via Attentional Mixtures of Soft Prompts	model efficiency (PEFT)	https://aclanthology.org/2022.emnlp-main.446
25	Parameter-Efficient Transfer Learning for NLP	model efficiency (PEFT)	https://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf
26	SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer	model efficiency (PEFT)	https://aclanthology.org/2022.acl-long.346
27	The Power of Scale for Parameter-Efficient Prompt Tuning	model efficiency (PEFT)	https://aclanthology.org/2021.emnlp-main.243
28	SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot	model efficiency (compression)	https://arxiv.org/pdf/2301.00774.pdf
29	An Efficient Memory-Augmented Transformer for Knowledge-Intensive NLP Tasks	model efficiency (inference)	https://arxiv.org/pdf/2210.16773.pdf
30	Hyperdecoders: Instance-specific decoders for multi-task NLP	model efficiency (PEFT)	https://arxiv.org/pdf/2203.08304.pdf
31	HyperPrompt: Prompt-based Task-Conditioning of Transformers	model efficiency (PEFT)	https://proceedings.mlr.press/v162/he22f/he22f.pdf

32	PERFECT: Prompt-free and Efficient Few-shot Learning with Language Models	model efficiency (training, inference)	https://arxiv.org/pdf/2204.01172.pdf
33	Learning to Compress Prompts with Gist Tokens	model efficiency (inference)	https://arxiv.org/abs/2304.08467