DataFlex: A Drop-In Upgrade That Makes LLM Training Data-Aware

Kabui, Charles

2026-04-06

Read at ToKnow.ai

Peking University's DCAI Lab released DataFlex, a unified framework built on LLaMA-Factory that brings dynamic data optimization into LLM training as a drop-in replacement. Instead of feeding all training data equally, DataFlex supports three strategies: sample selection (picking which examples matter most based on model state), domain mixture adjustment (shifting data ratios mid-training), and sample reweighting (assigning different importance to

individual examples). It unifies key operations like inference and gradient computation across all strategies, with support for multi-GPU distributed training at scale. In benchmarks, dynamic data selection consistently beat static full-data training on MMLU across Mistral-7B and Llama-3.2-3B, while two domain mixture methods (DoReMi and ODM) improved both accuracy and perplexity when pretraining Qwen2.5-1.5B on SlimPajama at 6B and 30B token scales.

Any team already using LLaMA-Factory can adopt DataFlex without rewriting training code: swap the trainer, add a few YAML config fields, and run. The framework pulls together scattered data optimization methods that previously lived in separate, often buggy codebases into one consistent interface. Researchers can now do apples-to-apples comparisons between strategies, and practitioners can test which approach works best for their dataset without juggling half a dozen repos.

The default in LLM training has long been "throw all data in with equal weight." DataFlex makes the alternative practical. As model architectures converge, how you schedule and weight your data during training may matter more than the architecture itself.

Sources:

- DataFlex Paper (arXiv)
- DataFlex GitHub (160 stars)
- DataFlex Documentation
- HuggingFace Daily Papers (#1, April 3)
