Uniflow Intro

Uniflow provides a unified LLM interface to extract and transform and raw documents. It enables data extraction from PDFs, HTMLs and TXTs, and supports most common-used LLMs for text transformation, including GPT4, Gemini 1.5, and Mistral-7B.

The Problems to Tackle

Uniflow addresses two key challenges in preparing LLM training data for ML scientists:

- first, extracting legacy documents like PDFs and Word files into clean text, which LLMs can learn from, is tricky due to complex PDF layouts and missing information during extraction; and
- second, the labor-intensive process of transforming extracted data into a format suitable for training LLMs, which involves creating datasets with both preferred and rejected answers for each question to support feedback-based learning techniques.

Hence, we built **Uniflow**, a unified LLM interface to extract and transform and raw documents.

Use Cases

Uniflow aims to help every data scientist generate their own private, ready-to-use training datasets for LLM finetuning, and hence make finetuning LLMs more accessible to everyone.

Here are some **Uniflow** hands-on solutions:

- Extract financial reports (PDFs) into summarization
- Extract financial reports (PDFs) and finetune financial LLMs
- Extract a math book (HTMLs) into your question answer dataset
- Extract PDFs into your question answer dataset
- Build RLHF/RLAIF preference datasets for LLM finetuning