

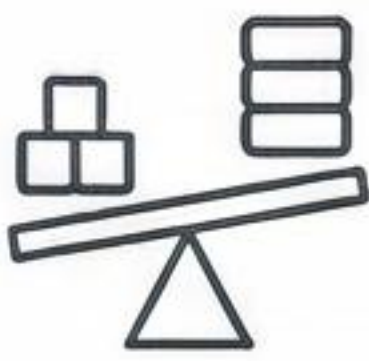
研究背景与挑战



• LLM数据准备流程碎片化，缺乏标准化。



• 现有框架对模型驱动的数据生成支持有限。



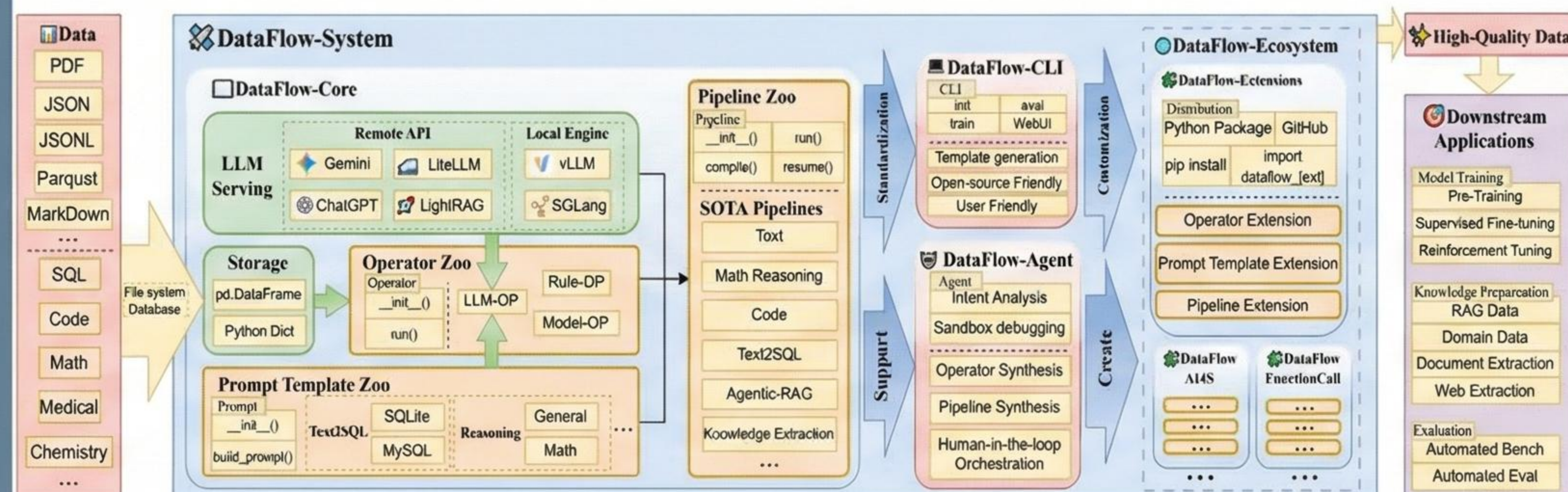
• 数据质量与多样性对模型性能至关重要。

现有框架对比 (Framework Comparison)

框架 (Framework)	特点 (Features)	局限性 (Limitations)
Framework A	Integrated pipeline	Limited generation
Framework B	Data versioning	Complex setup
Framework C	Model-in-loop	Scalability issues

DataFlow系统概述

- 目标：简化数据准备流程，提高可扩展性与自动化。
- 设计哲学：模块化、统一化、性能高效。
- 支持多领域数据准备与模型驱动工作流。



实验与结果分析

Table 2: 模型性能对比

任务类型	基线模型	新模型	提升
数学推理	-	+1-3分准确率	↑
Text-to-SQL	-	+3% 执行准确率	↑
代码	-	+7% 性能平均	↑
统一数据集	1M 样本	10K 样本	性能超越

- 数学推理数据：提升1-3分的准确率。
- Text-to-SQL数据：执行准确率提升+3%。
- 代码数据：平均提升7%的性能。
- 统一数据集：10K样本超越1M样本的性能。

研究贡献与意义

- 提出统一的LLM数据准备框架。
- 提供丰富的操作符与管道生态系统。
- 开发者友好的编程模型与开源支持。
- Agent驱动的自动化 workflows 构建。
- 实验验证数据质量与效率的提升。

框架的核心功能图

