

# DataFlow: 一个统一的LLM数据 准备与 workflow 自动化框架

作者: Hao Liang等

单位: 北京大学、上海人工智能实验室等

---

汇报人: XXX

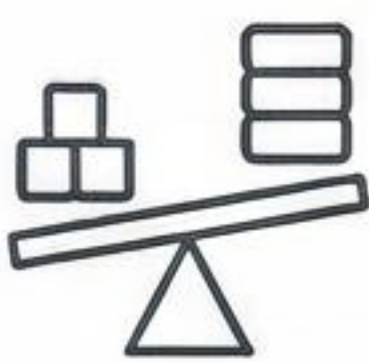
# 研究背景与挑战



• 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。



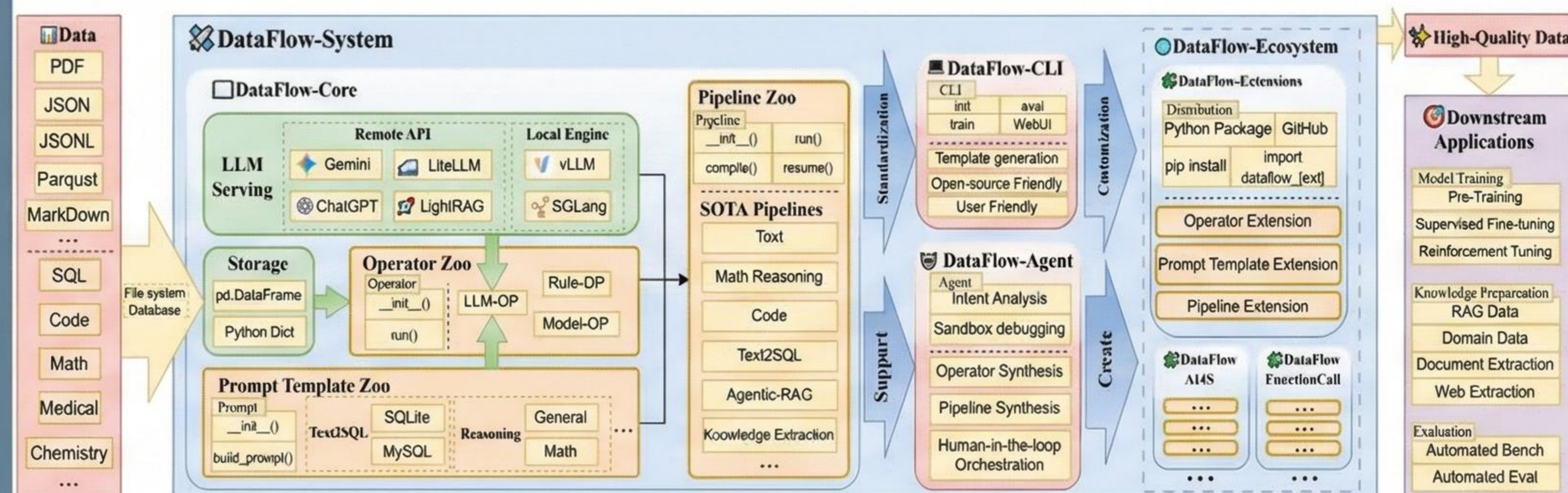
• 数据质量与多样性对模型性能至关重要。

现有框架对比 (Framework Comparison)

| 框架<br>(Framework) | 特点<br>(Features)    | 局限性<br>(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。



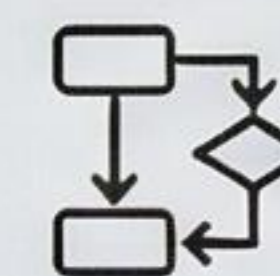
# 框架设计与架构

## 全局存储抽象



统一数据表示与存储。

## 操作符分类



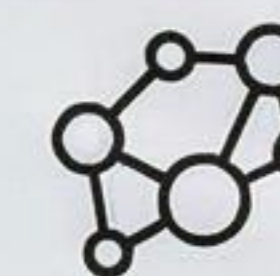
生成、评估、过滤与优化。

## 分层编程接口



支持LLM服务、操作符编程与模板接口。

## DataFlow生态系统

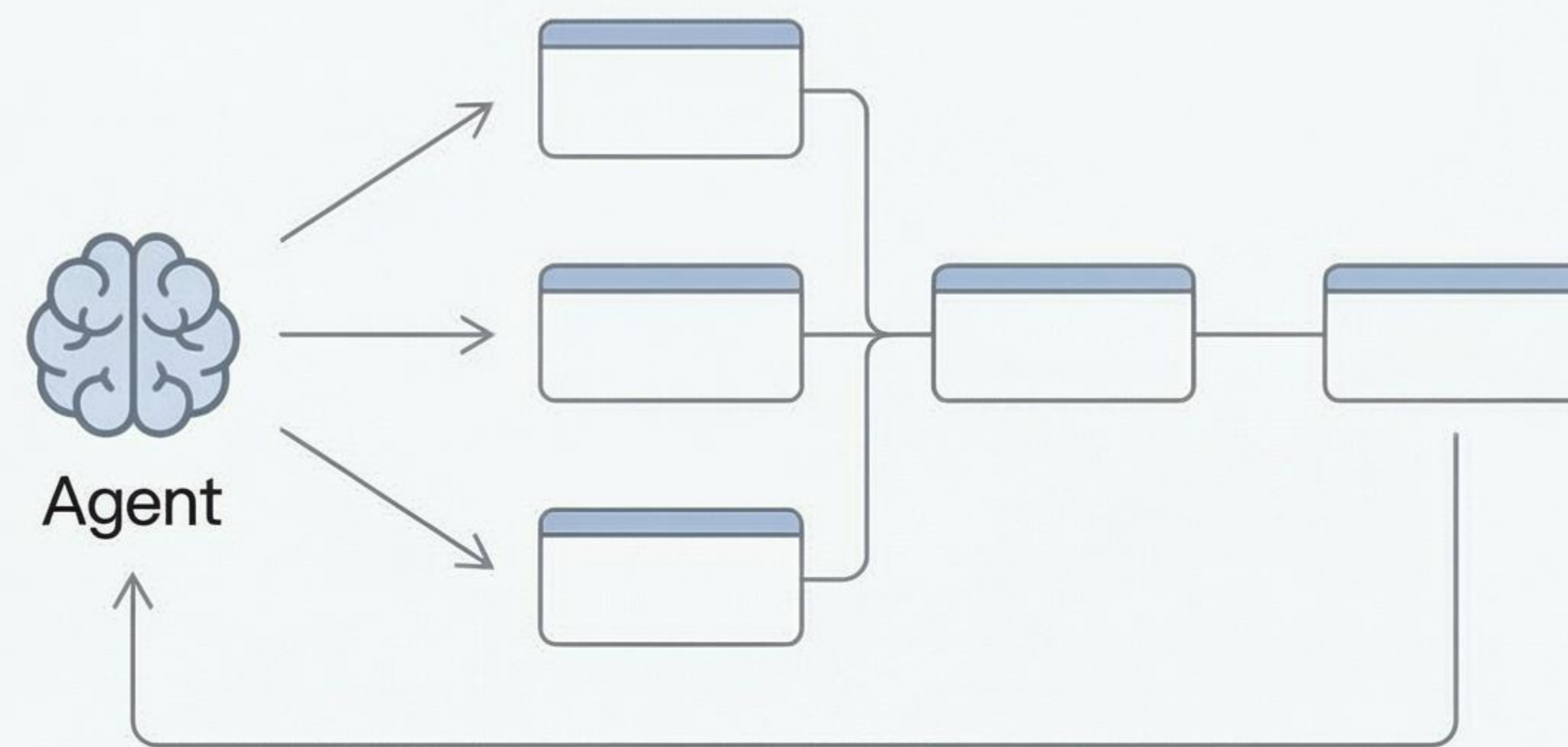


支持扩展与社区贡献。

# DataFlow-Agent与自动化 workflows

## Agent功能

- 从自然语言生成可执行工作流。
- 支持智能管道推荐与迭代验证。
- 降低构建复杂工作流的工程门槛。



迭代验证

工作流推荐示意图

# 实验与结果分析

Table 2: 模型性能对比

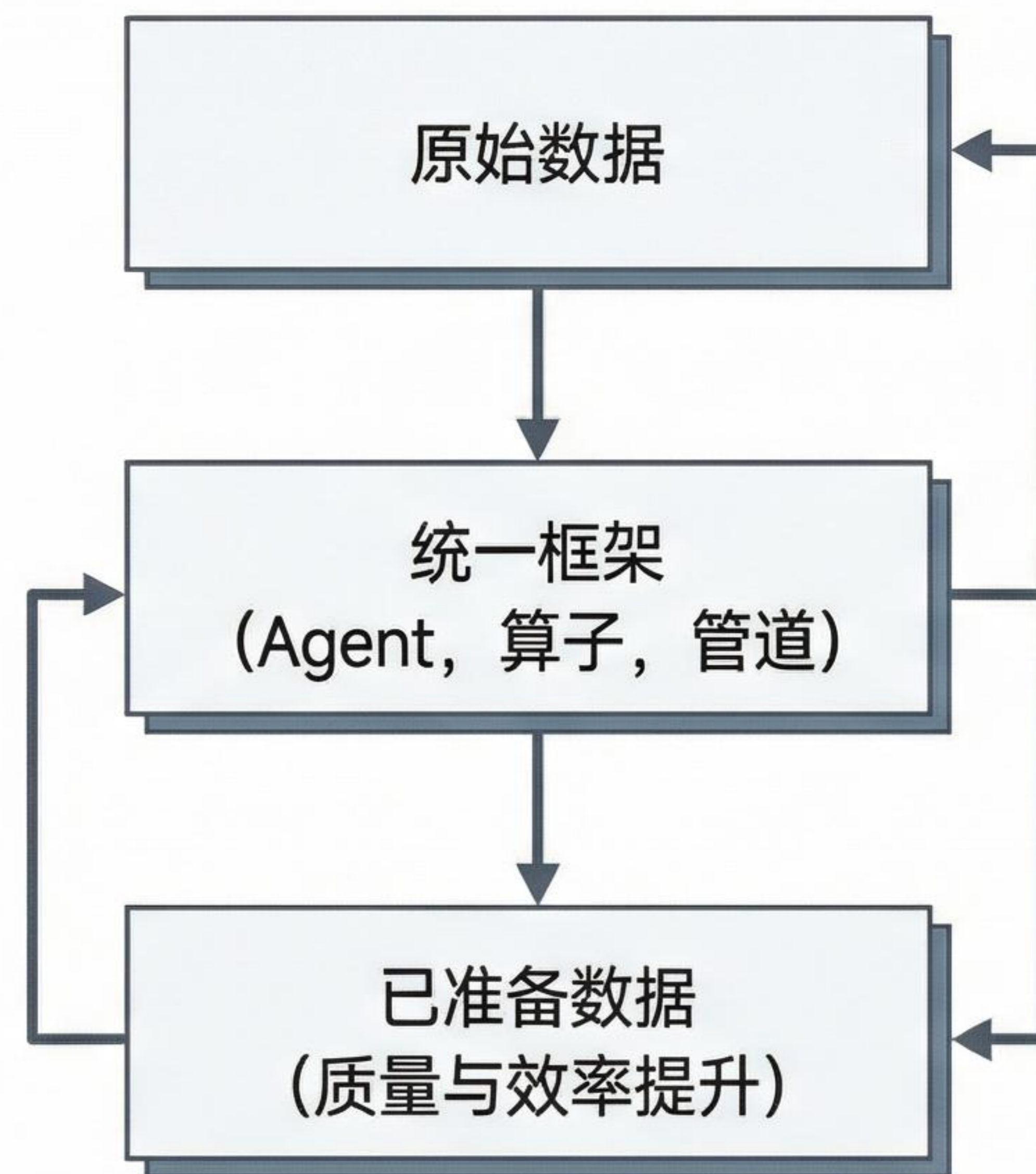
| 任务类型        | 基线模型  | 新模型       | 提升   |
|-------------|-------|-----------|------|
| 数学推理        | -     | +1-3分准确率  | ↑    |
| Text-to-SQL | -     | +3% 执行准确率 | ↑    |
| 代码          | -     | +7% 性能平均  | ↑    |
| 统一数据集       | 1M 样本 | 10K 样本    | 性能超越 |

- 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



# 致谢

---

- 感谢北京大学、上海人工智能实验室等单位的鼎力支持。
- 感谢所有作者与研究团队成员的辛勤付出与卓越贡献。
- 感谢相关资助项目与开源社区提供的宝贵资源与支持。

---

## 相关资助信息：

本研究工作得到了国家自然科学基金（项目批准号：No. XXXXXXXXX）及其他相关科技计划项目的资助。