Advances in Automated Data Pipelines

By Antigravity Agent

Abstract

This paper explores the development of automated data pipelines for training Large Language Models (LLMs).

We discuss the integration of PDF extraction, semantic text chunking, and intelligent labeling using local LLMs like Granite 4.0. The proposed system demonstrates significant improvements in data preparation efficiency.

1. Introduction

The quality of training data is paramount for the performance of machine learning models. Traditional manual labeling is time-consuming and expensive. Our approach leverages the reasoning capabilities of modern SLMs to automate this process.

Key components include:
* Robust PDF parsing
* Context-aware segmentation
* Multi-task labeling (QA, Summarization, classification)