GUI Wrapper for the DPK Planning Agent

Original Author: Dobromir Iliev (B.S. Math 2026)

Presented by - Kenneth (Alex) Jenkins (B.S. CompE 2026)

Georgia Tech

The Problem

- You’re playing around with some LLM training data – how do you know:
  - That your data isn’t poisoned?
  - Your data pipeline is okay?
  - Your sources are trustworthy?
  - Biases and anomalies are caught early?
  - The dataset is reproducible and well-documented?

👉 That’s where **IBM’s Data Prep Kit** comes in – giving you tools to validate, clean, and safeguard your training data before it ever reaches your model.

Georgia Tech

The Vision

- Georgia Tech Student and OSPO-VSIP intern, Dobromir Iliev, worked this summer to build a GUI for the Data Prep Kit (DPK) data transformations!
  - Using **Natural Language Processing (NLP)**, the GUI enables a feedback loop to design more optimal data pipelines with the right transformations.
  - This eliminates the need for separate Jupyter notebooks by leveraging the built-in **judge** and **planner** within the DPK repository.

GUI Wrapper for the DPK Planning Agent

Student: Dobromir Iliev
Mentors: Dr. Shahrokh Daijavad, Dr. Susan Malakia

Georgia Tech Open Source Program Office

Project Overview

- This project involves building an GUI for IBM's Data Prep Kit (DPK) for data transformations. It delivers a collection of tools for the creation, validation, and execution of data transformation pipelines, all accessible via both a robust CLI and an intuitive agentic GUI.
  - Automated Transformation & Validation: Provides a CLI and visual GUI for defining and executing data transformer pipelines

Goal

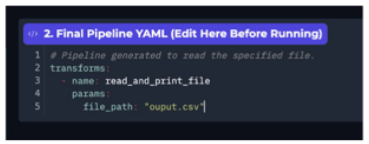
Understand the open-source project workflow while contributing a comprehensive CLI and agentic GUI wrapper.

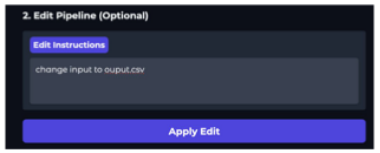
Milestones

- **Week 1-2**: Understand DPK and TTM model requirements. Set up the development environment with PyTorch, Hugging Face, Datasets, and DPK. Define CLI architecture and YAML schema for transformer chains.
- **Week 3-4**: Implement core CLI functionality for transformer sequences. Develop and integrate dataset validation logic. Prototype the agentic GUI, evaluating frameworks like Langflow and Gradio, and address initial integration.
- **Week 5-6**: Refine and complete the functional Gradio-based GUI. Integrate the CLI backend for visual pipeline construction and validation.
- **Week 7-11**: Finalize all code and comprehensive documentation for CLI and GUI. Create poster/slides for the final presentation, including DPK contributions.

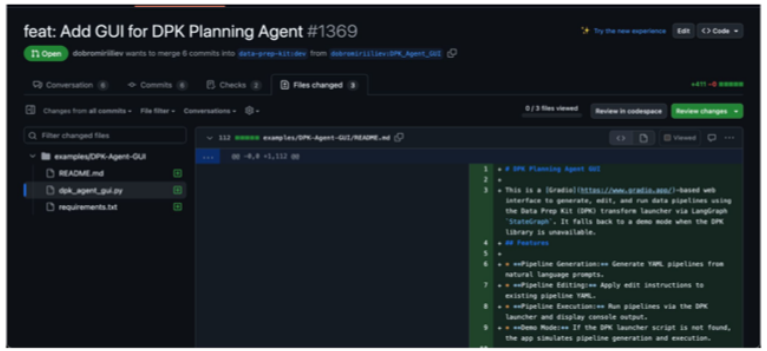
Highlights and Accomplishments

- The Final YAML Output; and the editable Pipelines





Highlights and Accomplishments



- Currently in the process of pushing the GUI onto the DPK repository

Learning Outcomes

- Gaining more familiarity with the open-source contribution workflow while navigating version control, specifically within the IBM Data Prep Kit (DPK) repository.
- Understanding the process of time series data transformers and validation logic.
- Getting hands-on experience in improving documentation for both CLI and GUI tools.

Future Work

- Further improve the documentation for both the CLI and Agentic GUI, including advanced usage and troubleshooting guides.
- Resolve any remaining small issues and perform bug fixes identified during testing and user feedback.
- Add new usage examples and demos for the CLI and GUI, showcasing diverse data transformation scenarios.
- Implement automated update/deployment features for the Dockerized toolchain, ensuring seamless distribution of new versions.

Georgia Tech

Georgia Tech