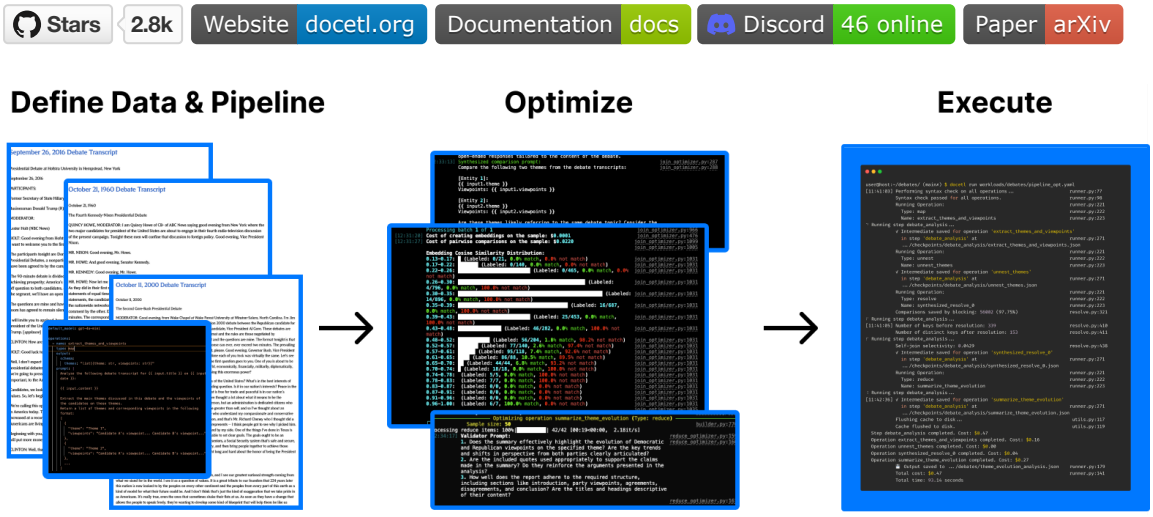
📜 DocETL: A System for Complex Document Processing



DocETL is a tool for creating and executing LLM-powered data processing pipelines. It offers a low-code, declarative YAML interface to define complex data operations on complex data.

When to Use DocETL

DocETL is the ideal choice when you're looking to **maximize correctness and output quality** for complex tasks over a collection of documents or unstructured datasets. You should consider using DocETL if:

- You have complex tasks that you want to represent via map-reduce (e.g., map over your documents, then group by the result of your map call & reduce)
- You're unsure how to best write your pipeline or sequence of operations to maximize LLM accuracy
- You're working with long documents that don't fit into a single prompt or are too lengthy for effective LLM reasoning
- You have validation criteria and want tasks to automatically retry when the validation fails

🚀 Features

- **Rich Suite of Operators**: Tailored for complex data processing, including specialized operators like "resolve" for entity resolution and "gather" for maintaining context when splitting documents.
- **Low-Code Interface**: Define your pipeline and prompts easily using YAML. You have 100% control over the prompts.
- **Flexible Processing**: Handle various document types and processing tasks across domains like law, medicine, and social sciences.
- **Accuracy Optimization**: Our optimizer leverages LLM agents to experiment with different logically-equivalent rewrites of your pipeline and automatically selects the most accurate version. This includes finding limits of how many documents to process in a single reduce operation before the accuracy plateaus.

⚡ Getting Started

To get started with DocETL:

1. Install the package (see installation for detailed instructions)
2. Define your pipeline in a YAML file. Want to use an LLM like ChatGPT or Claude to help you write your pipeline? See docetl.org/llms.txt for a big prompt you can copy paste into ChatGPT or Claude, before describing your task.
3. Run your pipeline using the DocETL command-line interface

🏛️ Project Origin

DocETL was created by members of the EPIC Data Lab and Data Systems and Foundations group at UC Berkeley. The EPIC (Effective Programming, Interaction, and Computation with Data) Lab focuses on developing low-code and no-code interfaces for data work, powered by next-generation predictive programming techniques. DocETL is one of the projects that emerged from our research efforts to streamline complex document processing tasks.

For more information about the labs and other projects, visit the EPIC Lab webpage and the Data Systems and Foundations webpage.