Simple Tutorial: Extracting Funny Quotes from Presidential Debates

In this tutorial, we'll walk through using the DocETL playground to extract funny or memorable quotes from presidential debate transcripts. We'll see how to:

1. Upload and explore data
2. Run a basic pipeline with sampling
3. Examine outputs
4. Iterate on prompts to improve results

Using LLMs to Help Write Pipelines

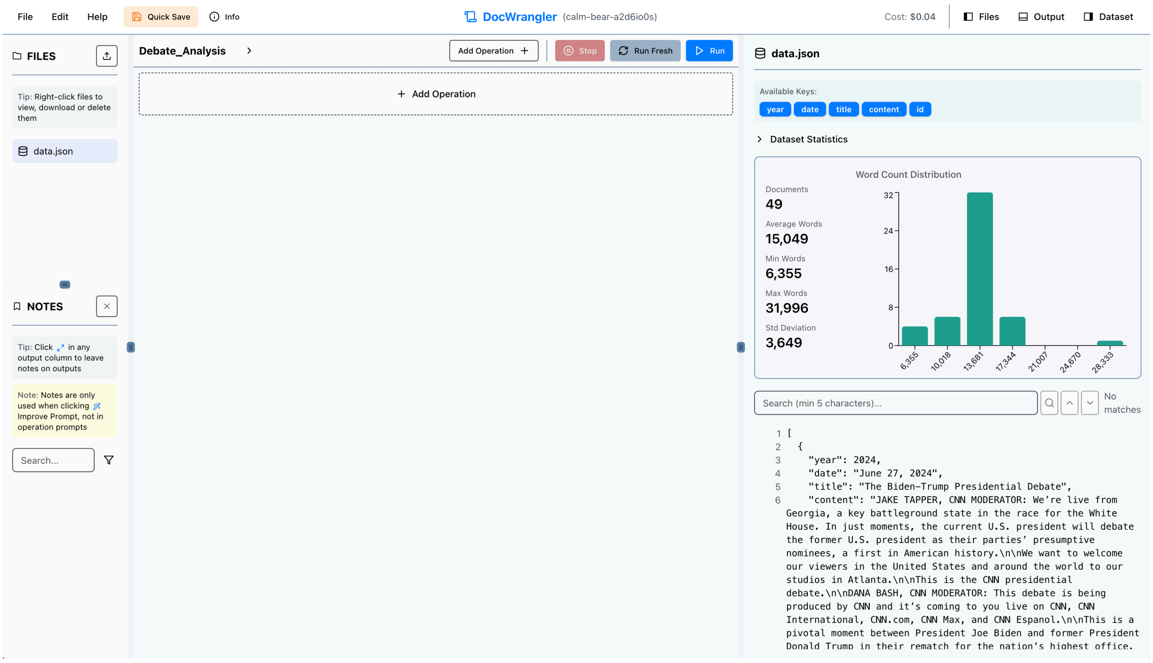
Want to use an LLM like ChatGPT or Claude to help you write your pipeline? See docetl.org/llms.txt for a big prompt you can copy paste into ChatGPT or Claude, before describing your task.

Step 1: Upload the Data

First, download the presidential debates dataset from here.

Once downloaded, use the left sidebar's upload button to load the data into the playground. The data contains transcripts from various presidential debates.

You can view the data by clicking the "toggle dataset view" button in the top right corner of the screen:



Tip: Right-click files to view, download or delete them

Tip: Click in any output column to leave notes on outputs

Note: Notes are only used when clicking Improve Prompt, not in operation prompts

Documents 49

Average Words 15,049

Min Words 6,355

Max Words 31,996

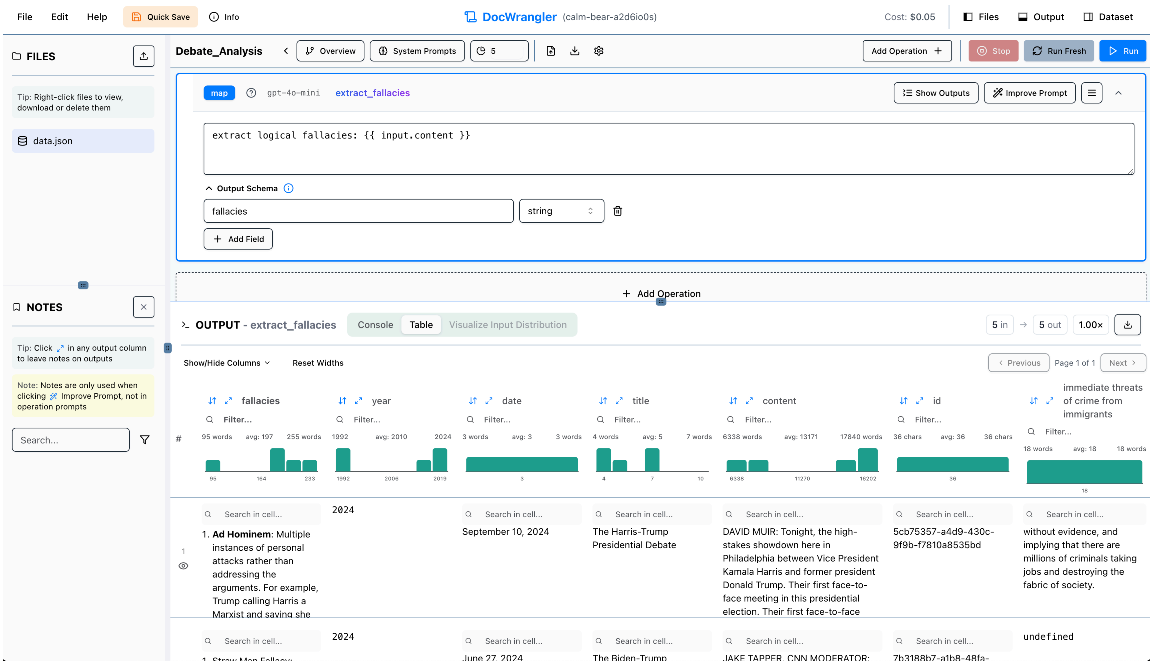
Std Deviation 3,649

```
[
  {
    "year": 2024,
    "date": "June 27, 2024",
    "title": "The Biden–Trump Presidential Debate",
    "content": "JAKE TAPPER, CNN MODERATOR: We’re live from Georgia, a key battleground state in the race for the White House. In just moments, the current U.S. president will debate the former U.S. president as their parties’ presumptive nominees, a first in American history.\n\nWe want to welcome our viewers in the United States and around the world to our studios in Atlanta.\n\nThis is the CNN presidential debate.\n\nDANA BASH, CNN MODERATOR: This debate is being produced by CNN and it’s coming to you live on CNN, CNN International, CNN.com, CNN Max, and CNN Espanol.\n\nThis is a pivotal moment between President Joe Biden and former President Donald Trump in their rematch for the nation’s highest office.
```

Step 2: Add a Map Operation

The pipeline is set to run on a sample of 5 documents, as indicated by the sample icon next to the pipeline name. This sampling helps us quickly test and iterate on our prompts without processing the entire dataset. This can be changed.

We'll add a map operation that processes each debate transcript to extract logical fallacies. Click the "Add Operation" button in the top right corner of the screen, and select "Map" under the LLM section. You can set the operation name to "extract_fallacies", and write a prompt and output schema.



Tip: Right-click files to view, download or delete them

Tip: Click in any output column to leave notes on outputs

Note: Notes are only used when clicking Improve Prompt, not in operation prompts

```
extract logical fallacies: {{ input.content }}
```

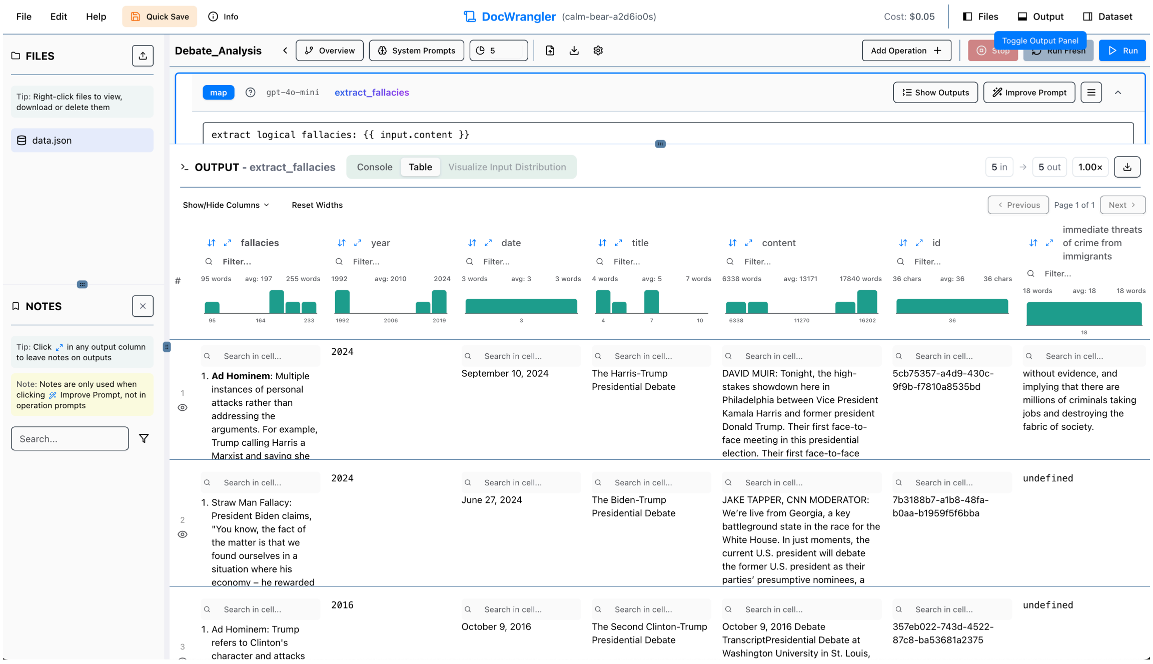
fallacies	string

#	fallacies	year	date	title	content	id	immediate threats of crime from immigrants
	95 words avg: 197 255 words	1992 avg: 2010 2024	3 words avg: 3 3 words	4 words avg: 5 7 words	6338 words avg: 13171 17840 words	36 chars avg: 36 36 chars	18 words avg: 18 18 words
1	1. **Ad Hominem**: Multiple instances of personal attacks rather than addressing the arguments. For example, Trump calling Harris a Marxist and saying she	2024	September 10, 2024	The Harris-Trump Presidential Debate	DAVID MUIR: Tonight, the high-stakes showdown here in Philadelphia between Vice President Kamala Harris and former president Donald Trump. Their first face-to-face meeting in this presidential election. Their first face-to-face	5cb75357-a4d9-430c-9f9b-f7810a8535bd	without evidence, and implying that there are millions of criminals taking jobs and destroying the fabric of society.
	1. Straw Man Fallacy:	2024	June 27, 2024	The Biden-Trump	JAKE TAPPER, CNN MODERATOR:	7b3188b7-a1b8-48fa-	undefined

Step 3: Run the Pipeline and Check Outputs

Click the "Run" button to execute the pipeline. The outputs panel will show two important tabs:

- **Console**: Displays progress information and any potential errors
- **Table**: Shows the extracted funny quotes from each document in a table, as well as the other key/value pairs in the document. Here's what the table looks like after running the pipeline:



```
extract logical fallacies: {{ input.content }}
```

#	fallacies	year	date	title	content	id	immediate threats of crime from immigrants
1	1. **Ad Hominem**: Multiple instances of personal attacks rather than addressing the arguments. For example, Trump calling Harris a Marxist and saying she	2024	September 10, 2024	The Harris-Trump Presidential Debate	DAVID MUIR: Tonight, the high-stakes showdown here in Philadelphia between Vice President Kamala Harris and former president Donald Trump. Their first face-to-face meeting in this presidential election. Their first face-to-face	5cb75357-a4d9-430c-9f9b-f7810a8535bd	without evidence, and implying that there are millions of criminals taking jobs and destroying the fabric of society.
2	1. Straw Man Fallacy: President Biden claims, "You know, the fact of the matter is that we found ourselves in a situation where his economy – he rewarded	2024	June 27, 2024	The Biden-Trump Presidential Debate	JAKE TAPPER, CNN MODERATOR: We're live from Georgia, a key battleground state in the race for the White House. In just moments, the current U.S. president will debate the former U.S. president as their parties' presumptive nominees, a	7b3188b7-a1b8-48fa-b0aa-b1959f5f6bba	undefined
3	1. Ad Hominem: Trump refers to Clinton's character and attacks	2016	October 9, 2016	The Second Clinton-Trump Presidential Debate	October 9, 2016 Debate TranscriptPresidential Debate at Washington University in St. Louis,	357eb022-743d-4522-87c8-ba53681a2375	undefined

You can resize the rows and columns in the table by clicking and dragging the edges of the table cells, as in the image above. You can also rezise the outputs panel by clicking and dragging the top edge of the panel.

Step 4: Iterate on the Prompt

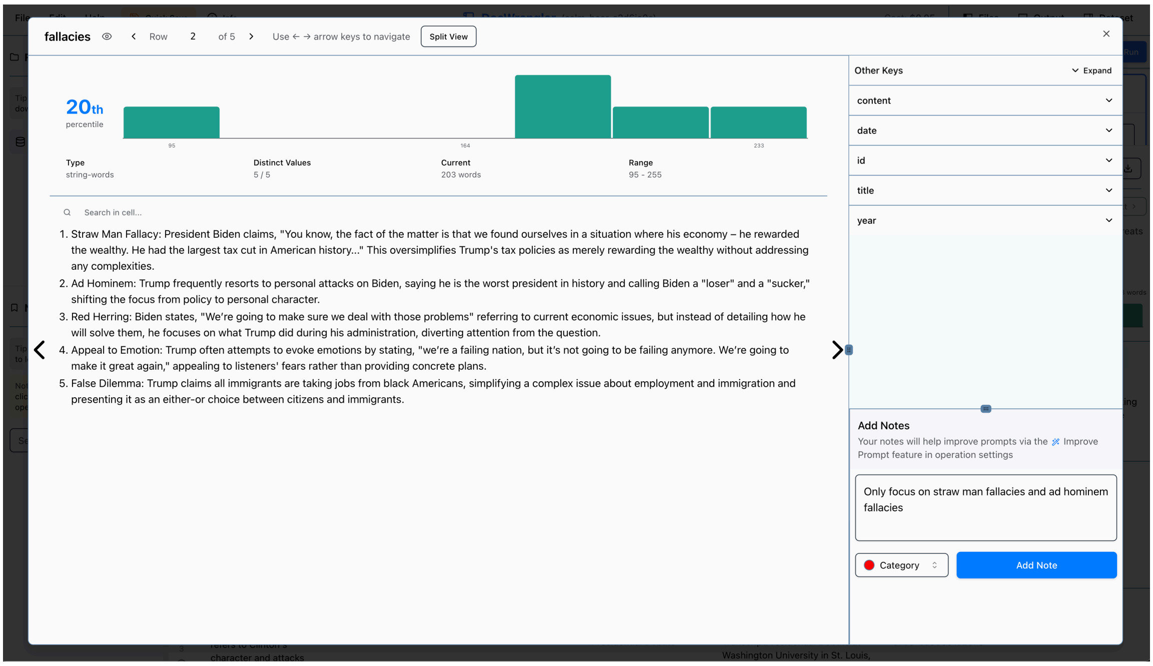
After reviewing the initial outputs, let's inspect them row by row to identify areas for improvement. Click the expand icon next to any column to see the full content.

We can modify the prompt to request additional context based on what we observe in the data. Here are some general patterns that could use improvement:

- Some quotes lack sufficient background context
- Speaker information could be more detailed
- The reasoning behind why quotes are memorable isn't always clear
- Time and setting details are often missing

- Impact and reactions to quotes aren't consistently captured

Here's an example of what giving notes might look like:



Type: string-words
Distinct Values: 5 / 5
Current: 203 words
Range: 95 - 255

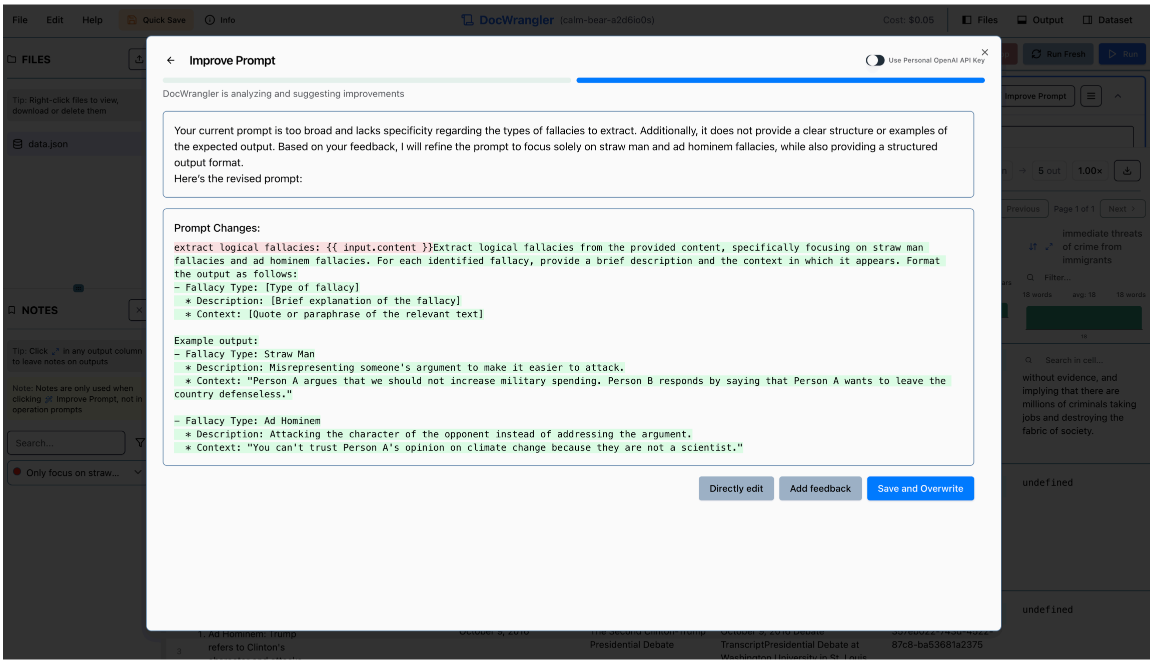
1. Straw Man Fallacy: President Biden claims, "You know, the fact of the matter is that we found ourselves in a situation where his economy – he rewarded the wealthy. He had the largest tax cut in American history..." This oversimplifies Trump's tax policies as merely rewarding the wealthy without addressing any complexities.
2. Ad Hominem: Trump frequently resorts to personal attacks on Biden, saying he is the worst president in history and calling Biden a "loser" and a "sucker," shifting the focus from policy to personal character.
3. Red Herring: Biden states, "We're going to make sure we deal with those problems" referring to current economic issues, but instead of detailing how he will solve them, he focuses on what Trump did during his administration, diverting attention from the question.
4. Appeal to Emotion: Trump often attempts to evoke emotions by stating, "we're a failing nation, but it's not going to be failing anymore. We're going to make it great again," appealing to listeners' fears rather than providing concrete plans.
5. False Dilemma: Trump claims all immigrants are taking jobs from black Americans, simplifying a complex issue about employment and immigration and presenting it as an either-or choice between citizens and immigrants.

Add Notes

Your notes will help improve prompts via the Improve Prompt feature in operation settings

Only focus on straw man fallacies and ad hominem fallacies

Once you've added enough notes (3-5 or more), you can click on the "Improve Prompt" button in the top right corner of the operation card. This will invoke the DocWrangler Prompt Improvement Assistant, which will suggest edits to your prompt:



DocWrangler is analyzing and suggesting improvements

Your current prompt is too broad and lacks specificity regarding the types of fallacies to extract. Additionally, it does not provide a clear structure or examples of the expected output. Based on your feedback, I will refine the prompt to focus solely on straw man and ad hominem fallacies, while also providing a structured output format.
Here's the revised prompt:

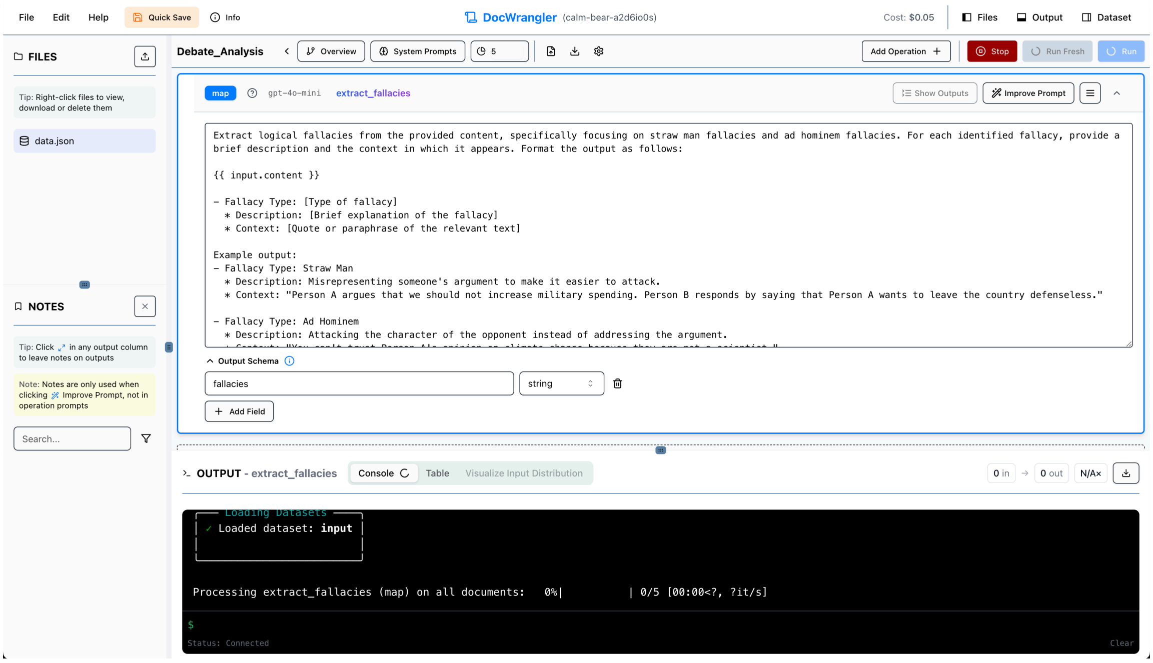
Prompt Changes:

```
extract logical fallacies: {{ input.content }}Extract logical fallacies from the provided content, specifically focusing on straw man fallacies and ad hominem fallacies. For each identified fallacy, provide a brief description and the context in which it appears. Format the output as follows:
- Fallacy Type: [Type of fallacy]
  * Description: [Brief explanation of the fallacy]
  * Context: [Quote or paraphrase of the relevant text]

Example output:
- Fallacy Type: Straw Man
  * Description: Misrepresenting someone's argument to make it easier to attack.
  * Context: "Person A argues that we should not increase military spending. Person B responds by saying that Person A wants to leave the country defenseless."

- Fallacy Type: Ad Hominem
  * Description: Attacking the character of the opponent instead of addressing the argument.
  * Context: "You can't trust Person A's opinion on climate change because they are not a scientist."
```

Once you're satisfied with the new prompt, click "Save" to update the operation, and then you can rerun the pipeline to see the new outputs.


File
Edit
Help
Quick Save
Info
DocWrangler (calm-bear-a2d6io0s)
Cost: $0.05
Files
Output
Dataset
FILES
Tip: Right-click files to view, download or delete them
data.json
NOTES
Tip: Click in any output column to leave notes on outputs
Note: Notes are only used when clicking Improve Prompt, not in operation prompts
Search...
Debate_Analysis
Overview
System Prompts
5
Add Operation
Stop
Run Fresh
Run
map
gpt-4o-mini
extract_fallacies
Show Outputs
Improve Prompt
Output Schema
Add Field
OUTPUT - extract_fallacies
Console
Table
Visualize Input Distribution
0 in
0 out
N/Ax
Clear

```
Extract logical fallacies from the provided content, specifically focusing on straw man fallacies and ad hominem fallacies. For each identified fallacy, provide a
brief description and the context in which it appears. Format the output as follows:

{{ input.content }}

- Fallacy Type: [Type of fallacy]
  * Description: [Brief explanation of the fallacy]
  * Context: [Quote or paraphrase of the relevant text]

Example output:
- Fallacy Type: Straw Man
  * Description: Misrepresenting someone's argument to make it easier to attack.
  * Context: "Person A argues that we should not increase military spending. Person B responds by saying that Person A wants to leave the country defenseless."

- Fallacy Type: Ad Hominem
  * Description: Attacking the character of the opponent instead of addressing the argument.
  * Context: "You can't trust Person A's opinion on climate change because they are not a scientist."
```

fallacies: string

```
┌── Loading Datasets ──┐
│ ✓ Loaded dataset: input │
│                         │
└─────────────────────────┘

Processing extract_fallacies (map) on all documents:   0%|          | 0/5 [00:00<?, ?it/s]

$
Status: Connected
```

Caching Behavior

DocETL automatically caches the outputs of each operation. This means that if you run the pipeline multiple times without changing the operations, it will use the cached results instead of reprocessing the documents. This is especially helpful when:

- Iterating on downstream operations in a multi-step pipeline
- Running the pipeline on the full dataset after testing on samples
- Sharing results with teammates (cached outputs persist across sessions)

The cache is invalidated only when you modify an operation's configuration (e.g., changing the prompt or schema). This ensures you always see fresh results when making changes while avoiding unnecessary recomputation.

If you want to bypass the cache and force recomputation, you can click the "Clear and Run" button instead of the regular "Run" button.

Step 5: Run the Pipeline on the Entire Dataset

Once you're satisfied with your prompt, you can run the pipeline on the entire dataset. First, clear the sample size by clicking on the settings or gear icon next to the pipeline name.

Then, click the "Run" button again. This will process all documents and update the outputs panel with the results.

You can export the results to a CSV file by clicking the "download" button in the top right corner of the outputs panel, near where it says "Selectivity: 1.00x". The selectivity is the ratio of the number of documents in the output to the number of documents in the input for that operation. In this case, since we ran the pipeline on the full dataset, the selectivity is 1.0x.

⚠ **Model Note**

In this tutorial, we used `azure/gpt-4o` instead of `gpt-4o-mini` since content filters were triggered by `gpt-4o-mini` when processing political debate content. If you encounter similar content filter issues with `gpt-4o-mini`, consider using `azure/gpt-4o` or another model with less restrictive filters.