

Enhancing Drupal Data Imports with LLMs

Jeremy Barth - <https://github.com/criolho/Drupal-Feeds-Using-LLMs>

Q. What is this talk about?

We will import 2 sources of Environmental Protection Agency (EPA) data into a custom Drupal site. We'll go beyond traditional web scraping by using an LLM to analyze and supplement the scraped data.

Q. Why would we want to do this?

PDFs and websites are valuable but **unstructured** (from our point of view) sources of data. Drupal, on the other hand, is all about **structured content**. LLMs help us add value over and above what the EPA provides: get better summaries, extract legal citations, get penalty info and assign categories based on our own taxonomy.

Q. How do we import the data and in what form?

- **Import** - there are many ways to bring content into Drupal, among others webhooks and the Migrate and Feeds modules. Feeds are a simple way to automate routine, scheduled imports and they have an intuitive UI.
- **Data format** - JSON is easy to read by both humans and machines. LLMs also like it.

Q. What are our data sources?

One site we'll be sourcing has an API while the other requires some web scraping.

Case Study #1 – Federal Register

The screenshot shows a web browser window displaying the Federal Register search results page. The browser's address bar shows the URL www.federalregister.gov/documents/search?conditions%5B. The page header includes navigation links for Sections, Browse, Search, Reader Aids, and My FR, along with a search bar for documents. The main content area features the Federal Register logo and the text "The Daily Journal of the United States Government". A blue bar highlights the "Document Search" section.

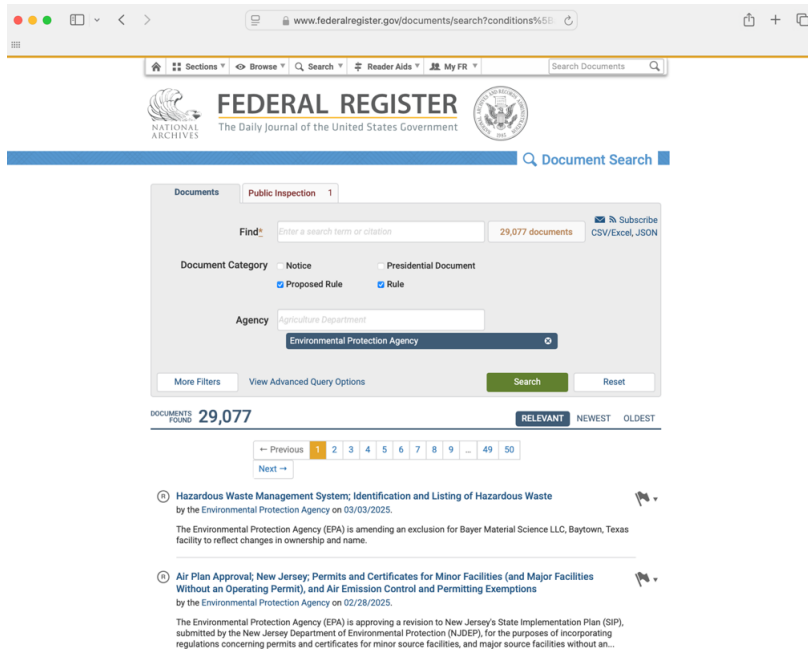
The search results are displayed in a grey box with the following filters and options:

- Documents:** Public Inspection 1
- Find:** Enter a search term or citation. 29,077 documents. [Subscribe](#) (CSV/Excel, JSON)
- Document Category:** Notice, Presidential Document, Proposed Rule, Rule
- Agency:** Agriculture Department (selected), Environmental Protection Agency (dropdown)
- Buttons:** More Filters, View Advanced Query Options, Search, Reset

Below the search box, the results are sorted by **RELEVANT** (options: RELEVANT, NEWEST, OLDEST). The total number of documents found is **29,077**. A pagination bar shows the current page is 1 of 50.

The first two search results are:

- Hazardous Waste Management System; Identification and Listing of Hazardous Waste** by the Environmental Protection Agency on 03/03/2025. The Environmental Protection Agency (EPA) is amending an exclusion for Bayer Material Science LLC, Baytown, Texas facility to reflect changes in ownership and name.
- Air Plan Approval; New Jersey; Permits and Certificates for Minor Facilities (and Major Facilities Without an Operating Permit), and Air Emission Control and Permitting Exemptions** by the Environmental Protection Agency on 02/28/2025. The Environmental Protection Agency (EPA) is approving a revision to New Jersey's State Implementation Plan (SIP), submitted by the New Jersey Department of Environmental Protection (NJDEP), for the purposes of incorporating regulations concerning permits and certificates for minor source facilities, and major source facilities without an...



Site: <https://federalregister.gov>

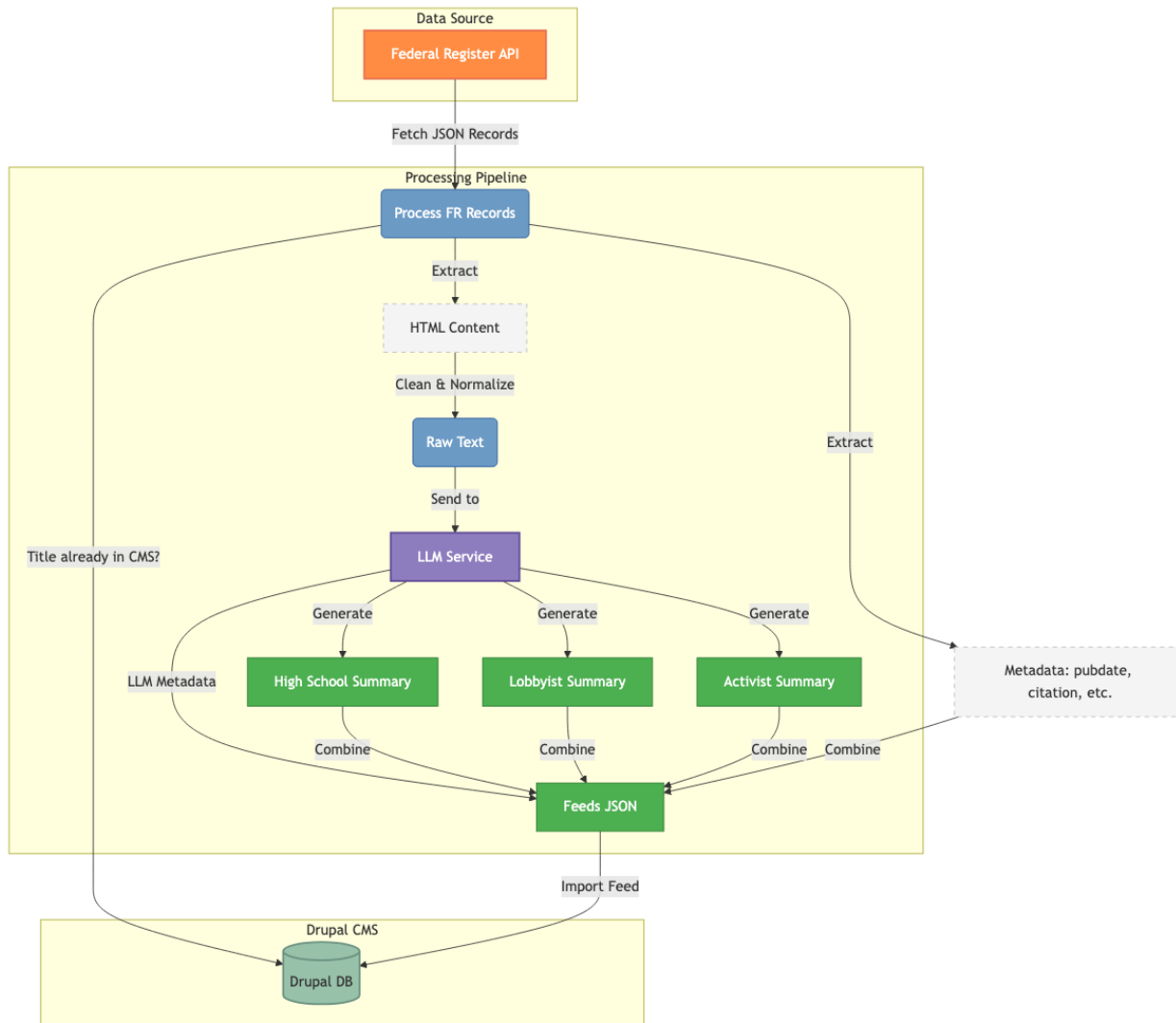
Background: This is where the U.S. government publishes all proposed federal regulations or rules about to go into effect. The site has an excellent UI and is also known for its developer-friendly API and JSON output.

Goals: connect to Federal Register API and get JSON metadata for recently-published rules. Then:

- Make minor transformations to their field data for our matching fields in Drupal, e.g. prepending a citation to their title to match our Drupal node title best practices
- Add 3 LLM-generated summaries individually focused on the needs of: 1) high school students, 2) corporate lobbyists, 3) environmental activists.

Federal Register processing pipeline

High level overview of the data processing game plan:



Inputs and Outputs

```
count: 5,  
description: "Documents published on or after 03/02/2025, from  
Environmental Protection Agency, and of type Rule or Proposed Rule",  
total_pages: 1,  
- results: [  
  - {  
    type: "Rule",  
    publication_date: "2025-03-07",  
    abstract: "The Environmental Protection Agency (EPA) is  
    approving a State Implementation Plan (SIP) revision  
    submitted through the South Carolina Department of Health  
    and Environmental Control (SC DHEC) on September 26, 2023,  
    regarding updates to the State's Cross-State Air Pollution  
    Rule (CSAPR) emissions trading programs. The SIP revision  
    incorporates by reference (IBRs) certain amendments EPA has  
    made to the regulations for the Federal CSAPR trading  
    programs for annual emissions of nitrogen oxides  
    (NO<INF>X</INF>) and sulfur dioxide (SO<INF>2</INF>) for  
    large electric generating units (EGUs). EPA is approving  
    South Carolina's September 26, 2023, SIP revision because it  
    is consistent with EPA's good neighbor CSAPR trading  
    programs and the Clean Air Act (CAA or Act).",  
    - agency_names: [  
      "Environmental Protection Agency"  
    ],  
    + cfr_references: [ ... ],  
    citation: "90 FR 11478",  
    effective_on: "2025-04-07",  
    document_number: "2025-03609",  
    pdf_url: "https://www.govinfo.gov/content/pkg/FR-2025-03-  
    07/pdf/2025-03609.pdf",  
    body_html_url:  
    "https://www.federalregister.gov/documents/full_text/html/20  
    25/03/07/2025-03609.html",  
    title: "Air Plan Approval; SC; Updates to the Cross-State  
    Air Pollution Rule",  
    + topics: [ ... ]  
  },  
  ]  
}
```

1. JSON from API

```
"documents": [  
  {  
    "type": "Rule",  
    "publication_date": "2025-03-07",  
    "abstract": "The Environmental Protection Agency (EPA) is approving a State Implementation Plan (SIP) ...",  
    "agency_names": "Environmental Protection Agency",  
    "cfr_references": "40 CFR 52",  
    "citation": "90 FR 11478",  
    "effective_on": "2025-04-07",  
    "document_number": "2025-03609",  
    "pdf_url": "https://www.govinfo.gov/content/pkg/FR-2025-03-07/pdf/2025-03609.pdf",  
    "body_html_url": "https://www.federalregister.gov/documents/full_text/html/2025/03/07/2025-03609.html",  
    "title": "90 FR 11478 - Air Plan Approval; SC; Updates to the Cross-State Air Pollution Rule",  
    "topics": "Air pollution control, Environmental protection, Incorporation by reference...",  
    "article_text": "AGENCY: Environmental Protection Agency (EPA). ACTION: Final rule. ...",  
    "high_school_summary": "...This plan is called the State Implementation Plan and it helps control pollution...",  
    "lobbyist_summary": "...This approval aligns with the Clean Air Act and involves incorporating by reference...",  
    "activist_summary": "...These pollutants contribute to environmental contamination, affecting ecosystems...",  
    "llm": "gpt-4o",  
    "ai_tags": "AI-Generated Text",  
    "time": "Fri, 07 Mar 2025 08:15:08"  
  }  
]
```

2. JSON for Drupal

Munged title (no AI)

Text fetch (no AI)

Summaries (AI-generated)

Data transformations

Case Study #2 – Civil Enforcement Actions by the EPA



- Environmental Topics
- Laws & Regulations
- Report a Violation
- About EPA

Home / Enforcement

- Enforcement Basics
- National Enforcement and Compliance Initiatives
- Enforcement and Compliance Assurance Annual Results for Fiscal Year 2024
- Air Enforcement
- Water Enforcement
- Waste, Chemical and Cleanup Enforcement
- Criminal Enforcement
- Enforcement at Federal Facilities
- Data and Results
- Policy, Guidance and Publications

[Enforcement: Contact Us](#)

Civil and Cleanup Enforcement Cases and Settlements

Currently available civil cases are listed below. Each case has a brief description and a link to detailed information about the case. You can filter the cases by statute or use the provided search box to search the table. You may also sort the data table by selecting a header.

[No Filter](#)
[BA](#)
[CAA](#)
[CAA112r](#)
[CERCLA](#)
[CWA](#)
[EPCRA](#)
[FIFRA](#)
[RCRA](#)
[TSCA](#)

Search:

Respondent	Description	Order Type	Date
Oasis Mobile Home Park Safe Drinking Water Act	On January 16, 2025, the U.S. Environmental Protection Agency (EPA) and the Department of Justice (DOJ) reached a settlement agreement with the operators of the Oasis Mobile Home Park (Park) for violations of the Safe Drinking Water at the Park (SDWA).	Consent Decree	2025-01-16
Stericycle, Inc. RCRA Settlement Summary	On January 16, 2025, the Environmental Protection Agency (EPA) and the Department of Justice announced a settlement agreement with Stericycle, Inc., a national provider of hazardous waste transportation, storage, and disposal services, for violations of the Resource Conservation and Recovery Act's (RCRA) regulations regarding hazardous waste transport, storage, and recordkeeping.	Stipulation and Order of Settlement and Complaint	2025-01-16
	On January 16, 2025, the Environmental Protection		

The screenshot shows the EPA website's 'Civil and Cleanup Enforcement Cases and Settlements' page. The page features a navigation menu on the left, a search bar at the top, and a table of cases. The table has columns for Respondent, Description, Order Type, and Date. Two cases are visible: one involving Oasis Mobile Home Park and another involving Stericycle, Inc.

Respondent	Description	Order Type	Date
Oasis Mobile Home Park, Safe Drinking Water Act	On January 16, 2025, the U.S. Environmental Protection Agency (EPA) and the Department of Justice (DOJ) reached a settlement agreement with the operators of the Oasis Mobile Home Park (Park) for violations of the Safe Drinking Water at the Park (SDWA).	Consent Decree	2025-01-16
Stericycle, Inc., RCRA Settlement Summary	On January 16, 2025, the Environmental Protection Agency (EPA) and the Department of Justice announced a settlement agreement with Stericycle, Inc., a national provider of hazardous waste transportation, storage, and disposal services, for violations of the Resource Conservation and Recovery Act's (RCRA) regulations regarding hazardous waste transport, storage, and recordkeeping.	Stipulation and Order of Settlement and Complaint	2025-01-16
	On January 16, 2025, the Environmental Protection		

Site: <https://www.epa.gov/enforcement/civil-and-cleanup-enforcement-cases-and-settlements>

Background: Centralized source for EPA civil actions for violations. The EPA [historically] enforces a wide range of environmental laws, including the Clean Air Act (CAA) and Clean Water Act (CWA).

Goals:

- Better summaries
- Extract penalty info
- What specific laws were violated?
- Categorize environmental issues

Inputs and Outputs

Turn 14 Clean Air Act Settlement Summary

On January 17, 2025, the Environmental Protection Agency (EPA) and Department of Justice reached a settlement agreement with Turn 14 Distribution, Inc. ("Turn 14"), for violations of the Clean Air Act from the sale of devices that "defeat" emissions control systems in cars and trucks.

Defeating vehicle emission controls causes the release of excess air pollution including nitrogen oxides (NOx), nonmethane hydrocarbons (NMHC), carbon monoxide (CO), and particulate matter (PM) above legal limits. Excess pollution from vehicles with defeat devices harms public health and impedes efforts by EPA, tribes, states, and local agencies to plan for and attain air quality standards. Under the settlement, the defendant will pay \$3.6 million.

On this page:

- [Overview of Turn 14](#)
- [Summary of the Violations](#)
- [Summary of Environmental and Health Impacts](#)
- [Overview of the Consent Decree](#)
- [Comment Period](#)
- [Contact Information](#)

Settlement Resources

- [Turn 14 Consent Decree \(pdf\)](#) (4.94 MB)
- [Turn 14 Complaint \(pdf\)](#) (304.24 KB)

Overview of Turn 14

Turn 14 is one of the country's largest wholesale distributors of automotive parts and suppliers in the United States, with warehouses located in Hatfield, PA, Reno, NV, Arlington, TX, and Indianapolis, IN. Turn 14 purchases products from manufacturers and sells them to other distributors or retail outlets through its private, online, business-to-business website.

Summary of the Violations

Between January 2016 and August 2021, Turn 14 sold over 140,000 of these tampering devices throughout the United States.

Fetch details page

Respondent	Description	Order Type	Date
Turn 14 Clean Air Act Settlement Summary	On January 17, 2025, the Environmental Protection Agency (EPA) and Department of Justice reached a settlement agreement with Turn 14 Distribution, Inc. ("Turn 14"), for violations of the Clean Air Act from the sale of devices that "defeat" emissions control systems in cars and trucks.	Consent Decree and Complaint	2025-01-17

JSON for Drupal Feed

```
{
  "documents": [
    {
      "order_type": "Consent Decree and Complaint",
      "date": "2025-01-17",
      "link": "https://www.epa.gov/enforcement/turn-14-clean-air-act-settlement-summary",
      "pdf_links": "https://www.epa.gov/system/files/documents/2025-03/turn-14-dn-2-notice-of-lodging-c...w-prop-cd.pdf,https://www.epa.gov/system/files/documents/2025-03/turn-14-dn-1-complaint.pdf",
      "title": "EPA Enforcement Turn 14 Clean Air Act Settlement Summary",
      "raw_text": "Turn 14 Clean Air Act Settlement Summary On January 17, 2025, the Environmental Protection Agency (EPA) and Department of Justice reached a settlement agreement with Turn 14 Distribution, Inc. (\"Turn 14\"), for violations of the Clean Air Act from the sale of devices that \"defeat\" emissions control systems in cars and trucks. Defeating vehicle emission controls causes the release of excess air pollution including nitrogen oxides (NOx), nonmethane hydrocarbons (NMHC), carbon monoxide (CO), and particulate matter (PM) above legal limits. Excess pollution from vehicles with defeat devices harms public health and impedes efforts by EPA, tribes, states, and local agencies to plan for and attain air quality standards. Under the settlement, the defendant will pay $3.6 million."
    },
    {
      "summary": "<p>On January 17, 2025, the <b>Environmental Protection Agency (EPA)</b> and <b>Department of Justice</b> reached a settlement agreement with <b>Turn 14 Distribution, Inc.</b>, for violations of the <b>Clean Air Act</b>. The violations stemmed from the sale of devices that \"defeat\" emissions control systems in cars and trucks. These \"defeat devices\" cause the release of excess air pollution, including nitrogen oxides (NOx), nonmethane hydrocarbons (NMHC), carbon monoxide (CO), and particulate matter (PM) above legal limits.</p><p>Between January 2016 and August 2021, <b>Turn 14</b> sold over 140,000 of these tampering devices throughout the United States, violating Section 203(a)(3)(B) of the <b>Clean Air Act</b>. Car, truck, and engine manufacturers install emission controls on vehicles and engines to comply with <b>Clean Air Act</b> emission standards. <b>Turn 14</b> sold devices designed to remove or defeat emission controls from numerous vehicle models.....\"
    },
    {
      "penalty": "3600000.0",
      "environmental_issues": [
        "Automobiles and Trucks"
      ],
      "laws": {
        "federal_law": [
          { "type": "Statute", "citation": "42 U.S.C. § 7522(a)(3)(B)" },
          { "type": "Rule", "citation": "28 C.F.R. § 50.7" },
          { "type": "Rule", "citation": "40 C.F.R. § 86.1803-01" }
        ],
        "flattened_federal_laws": "Statute - 42 U.S.C. § 7522(a)(3)(B),Rule - 28 C.F.R. § 50.7,Rule - 40 C.F.R. § 86.1803-01"
      },
      "llm": "models/gemini-2.0-flash",
      "time": "Sat, 08 Mar 2025 12:08:26",
    }
  ]
}
```

1. From web page

2. From LLM: Summary, Penalty, Issues, Laws

LLM considerations

Closed models, open source – they’re constantly leapfrogging one another. At present, closed source foundation models produce more **reliable structured output**. This may change in the future.

The code here has options for OpenAI’s “gpt-4o”, Anthropic’s "claude-3-7-sonnet-latest" and Google’s “gemini-2.0-flash”. Currently we default to **Gemini Flash** because it is fast, reliable, inexpensive (as of March 2025: input \$0.10 / 1,000,000 input tokens, \$0.40 / 1,000,000 output tokens) and has a long context window (1,000,000 tokens) that can readily accommodate multi-hundred page PDFs.

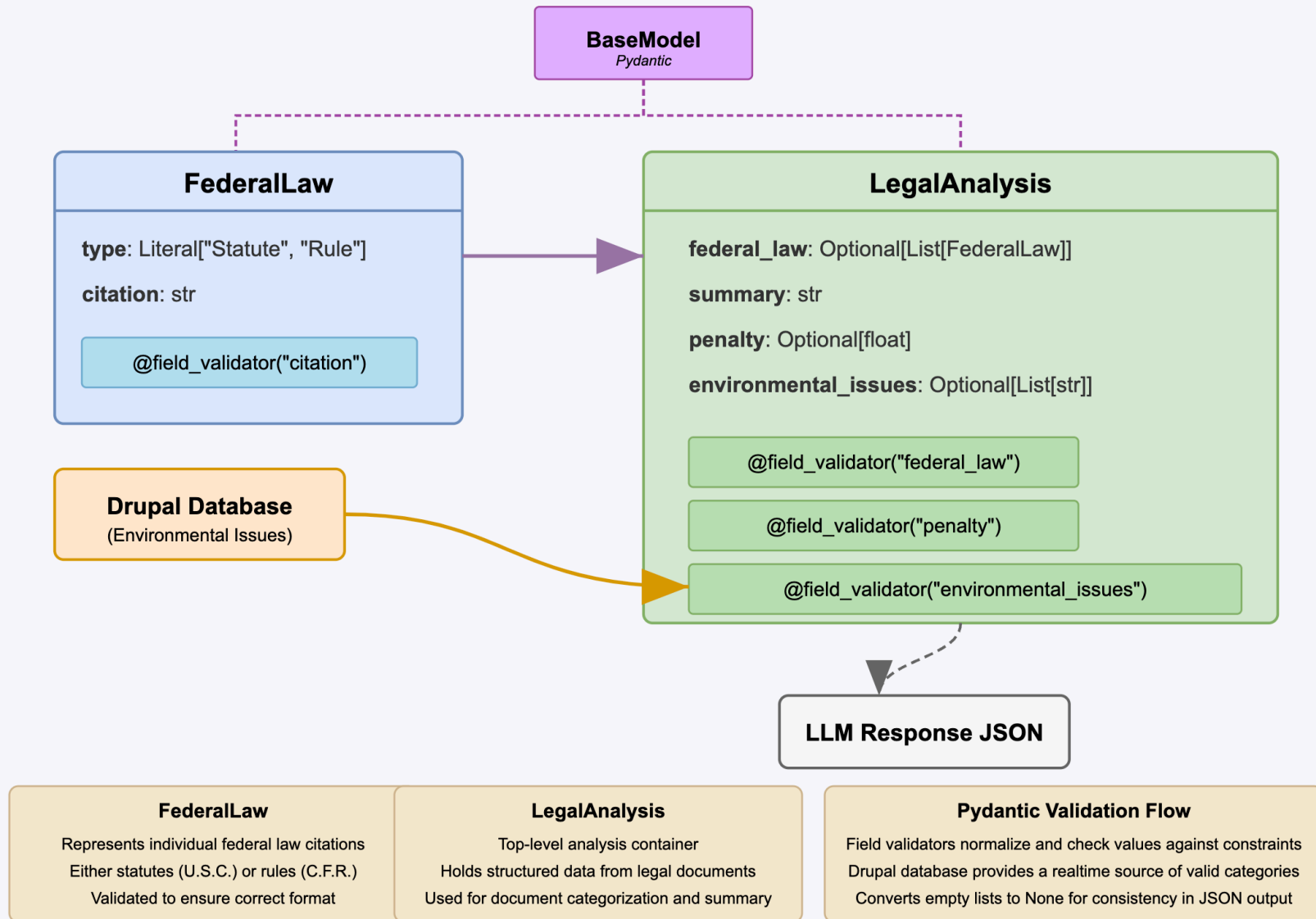
Optimization – a custom `title_exists()` function checks to see if we already have a node in Drupal, saving on unnecessary LLM calls

Pydantic and Instructor

Pydantic and Instructor are Python libraries that work together to: a) tell LLMs how you want their reply to be structured, e.g. JSON with certain fields and types of data; b) validate the data before passing it on for downstream use. In particular, we want anything GenAI-related to pass muster before we import it to our Drupal CMS.

You define a “data model” and Pydantic helps ensure that it’s followed, e.g. that certain fields must be present or are merely optional. You can also defined custom validation methods.

Pydantic Data Models



Pydantic Field validators

These are optional functions you can write that automatically run against fields before Pydantic signs off on the data. You can use them to simply confirm that data matches certain criteria, or you can modify the data to conform your needs. The point is that this is a **structured way of enforcing standards** for your data.

What we're choosing to validate:

- **citation** – make sure legal citations are in a standardized format, e.g. 40 C.F.R. §§ Part 1039" should be transformed to "40 C.F.R. § 1039
- **penalty** – make sure it's a numeric float with at most 2 decimal places
- **environmental_issues** – suppose we have a Drupal taxonomy we're already using. We can fetch the terms dynamically from the Drupal DB and use them both to tell the LLM what terms we want it to look for and then to make sure that's what it did.

Miscellaneous “Best Practices”

There are few “standards” yet for how websites should manage AI-generated content. You may want to consider:

- Creating a vocabulary “AI Tags” to help you keep track of nodes to which you’ve applied GenAI. For example: AI-Generated Text, AI-Generated Categories, or AI-Generated Entity Extraction
- It is quite likely that you’ll use different LLMs over time. You may want to have a vocabulary for these as well with terms such as: claude-3-7-sonnet-latest, gpt-4o, gpt-4o-mini, gemini-2.0-flash
- If you’re going to the trouble of extracting lots of raw text for use in GenAI, even though in its raw state it may not be suitable for end users it might be good for a) fulltext search; b) future GenAI passes over the same nodes
- People are justifiably wary of what’s being pushed on them – consider including preamble / info text such as “This article contains an AI-generated summary” that fully informs people what they’re getting.