



GenAI 301: AI + APIs

Full Team Call

October 2025

Objectives

After this session, you should walk away understanding:

1. What an API is
2. How it differs from other ways of interacting with GenAI models
3. When you might choose to use an API versus the chat interface
4. What using an API with a GenAI model looks like

As part of this, we will also walk through the recent media analysis Bellwether did using OpenAI's API.

Agenda

Introduction to APIs	20 mins
Example: Media Analysis	25 mins
Q&A	15 mins



Introduction to APIs

What is an API?

- API stands for **A**pplication **P**rogramming **I**nterface
- APIs **allow two pieces of software to “talk” to each other and share data directly**
- Instead of a human manually transferring information (e.g., USB drive upload or copy-pasting), **APIs use programming language (aka code) to request and deliver information.**

A good metaphor for an API is a restaurant, where the diner is the user, the waiter is an API, and the kitchen/chef is the other side.



You (the user) aren't bringing in your own ingredients (information) to be cooked.

Instead, you **request** certain foods via the waiter (API), who brings that request (along with any relevant information like allergies) to the kitchen.

The kitchen prepares a dish according to your request, and the waiter (API) brings it back out to you.

APIs have been around for a long time, and power many of our everyday experiences

- APIs have existed for decades as the “connective tissue” among digital systems
- You likely send and receive information via APIs everyday:



Getting real-time weather forecasts on your phone requires your weather app to interface with weather systems using an API



Sharing a photo or video from one app to another (e.g., Instagram → Facebook) uses an API so you don't have to re-upload or copy-paste



Smart home integrations (e.g., “Alexa, turn off the lights”) use APIs to translate and execute your commands



Online payment or POS systems use APIs to transfer financial information (usually encrypted)

- Note: although they use code, APIs aren't just for software developers
 - Many people set up API connections through “no-code” tools like Zapier, IFTTT, etc. to make apps work better for their specific context or use case

!! APIs exist and work **completely independent of GenAI**. We can use APIs to connect to GenAI models, but APIs can be used with *any* software, not just AI-powered apps.

So how *do* APIs work with GenAI?

- Typically, users “talk” to a GenAI model through a **chatbot user interface**
- This allows us to send requests and get responses in a language easy to understand
- However, the **chat interface can be limited for certain use cases:**
 - Repetitive tasks across a LOT of data (think hundreds or thousands of rows)
 - Sending very large files (100+ page documents, long video or audio files)
 - Building custom applications
- An API connects directly to the model (no chat interface) to address these issues:
 - Code can **automatically send repeat prompts** to a model and aggregate the outputs
 - A backend API connection allows developers to create **custom front-end user interfaces** that only show what is necessary for their intended users

Essentially, APIs let us move from “chatting” with a model to *integrating* GenAI into our own tools and workflows.



In practice, an example could be an edtech tool using APIs with a GenAI model to:

1. Pull student test scores from the school’s SIS
2. Send those to a GenAI model for analysis
3. Take the output from the GenAI and reformat it into a dashboard that displays trends or key themes

A man with short dark hair is looking down at a computer monitor. The screen displays various data visualizations, including several bar charts and a line graph. He is wearing a light-colored button-down shirt. In the background, there are more blurred screens showing similar data, suggesting a control room or monitoring station.

A Real Example: Media Analysis

First, some grounding: ICYMI...

Myself, Nora, Katrina, Julie, and Andy J. analyzed more than 1,500 articles from national news outlets to understand:

- What are the **dominant themes in national media coverage** of K-12 education?
- Who are the **most frequently quoted organizations and individuals** in education reporting, and how do they shape narratives?
- To what extent **does media coverage prioritize research, policy solutions, and actionable insights** around education issues, versus opinion or commentary?
- **How have trends on education coverage changed** over the course of the past year, especially in relation to major national events such as the 2024 presidential election, inauguration, and changes at the U.S. Department of Education?



A Year in Review: How National Media Covers K-12 education

SEPTEMBER 24, 2025

NORA WEBER, MARISA MISSION, KATRINA BOONE, JULIET SQUIRE, AND ANDY JACOB

Check out the final product [here!](#)

This was our first AI-assisted, public-facing product

As a result, we were *very meticulous* about the methodology – when you get a chance, please take a read to better understand how we’re thinking about **rigorous AI usage!**

Ensuring Rigor When Using AI

The use of machine learning and natural language processing is neither new nor uncommon in content analyses.¹ The use of newer generative AI and large language models (LLMs), on the other hand, has undergone less testing — but research suggests promise in its application.² Using an LLM for content analyses can greatly improve speed and efficiency, allowing for deeper analysis. At the same time, it also introduces new risks, such as the potential for fabricated data, inaccurate coding, and a lack of transparency due to the “black box” of how these models function.³ To mitigate these risks, the Bellwether team grounded this analysis in validated design and analytical practices, as well as consulted resources and used the following strategies to test the reliability and validity of results.

Strategy 1: Guardrails for Validity

These validity criteria were determined prior to beginning the analysis and set safeguards to ensure that analysis did not move forward on faulty assumptions:

- A minimum of 10% of the sample, chosen at random, must be first coded by humans independent of one another and independent of seeing the AI results.
- The thematic analysis codebook must be developed by humans independent of AI.
- Prompts must be developed through iterative and progressive testing, where AI outputs are checked against the human coding. The analysis will not move forward unless the results reach a minimum threshold of 80% accuracy.
- After the analyses are run on the full sample of articles, humans will conduct spot checks of the AI results.
- All prompts, model versions, validation sets, thresholds, and codebook changes must be documented.

Strategy 2: Using Retrieval for Thematic Analysis Accuracy



We grounded our approach with a heavy emphasis on “humans in the loop,” with AI just scaling our own thinking

Phase of Research	Phase 1: Extraction of Deductive Metadata	Phase 2: Thematic Coding and Extraction of Inductive Attributes	Phase 3: Trend Analysis
Human Design, Validation, and Analysis	<ul style="list-style-type: none"> Cleaned dataset of online articles. Identified key article themes by manually coding 50 articles and validating results with two separate analysts, then compiled those themes into a codebook for consistency. Created a prompt for OpenAI API to extract key article details (e.g., author, date, outlet). Iterated on AI prompt until OpenAI results aligned with human results at a minimum of 80% accuracy, based on a comparison of results in Microsoft Excel: <ul style="list-style-type: none"> Tested prompt on the manually-coded sample. Validated results against manual coding. Refined prompt, as needed. 	<ul style="list-style-type: none"> Manually coded 100 additional articles for a total of 150 manually-coded articles. Created a thematic analysis prompt to use with the OpenAI API. Iterated on the prompt until results reached a minimum of 80% alignment with human results, based on a comparison of results in Microsoft Excel: <ul style="list-style-type: none"> Tested prompt on the 150 manually-coded articles. Validated results against manual coding. Refined prompt, as needed. 	<ul style="list-style-type: none"> Spot-checked the final deductive and inductive results from the AI-powered coding. Removed articles that were duplicates or not related to education (e.g., special interest stories). Analyzed the final dataset according to project's research questions.
AI Tool	<ul style="list-style-type: none"> Pulled key details (e.g., author, date, outlet) from every article in the sample to create a dataset of article attributes. 	<ul style="list-style-type: none"> Converted the articles and theme codebook into numeric representations and stored them in a specialized search index to be used for thematic analysis. Conducted the thematic analysis by identifying each article's main topic and coding it according to the theme codebook. 	<ul style="list-style-type: none"> N/A



Now onto the good stuff!

Step 1: Deciding to Use the API vs Chat

Several indicators that the team should use the API and not the chat interface:

1. Volume of content to be analyzed:

We received nearly 1600 articles in two packages of 800+ page Word documents 😱

2. Repetitive quality of the analysis:

Each article varied in its contents, but the analysis itself (extracting metadata and coding it to a theme) would stay the same across every article

3. High-quality but inconsistent responses from chat interface:

In tests, various models of ChatGPT could complete the analysis on 5-10 articles, but not on an entire article package, signaling that the models had the *capability*, but not the *capacity*

Extracted Articles Metadata					
	Title	News Outlet	Date	People Quoted	Data Sources
1	National Education Media Coverage	April 1, 2023 - April 24, 2024	April 1, 2023		
2	USA Today		April 29, 2024		
3	Fox	Opinion: The state of the American teenager	April 29, 2024		reports of students saying that they are unsafe in school, let alone 43% of teens feeling unsafe;
4	CBS	Technology education program wants to bring out the geniuses in Chicago's young	April 28, 2024		reporters helped USA TODAY identify roughly 370 campuses based in nearly 200
5	Fox		April 23, 2024	Dade County; President Biden	report emergencies to law enforcement and has an anonymous tip reporting function
6	Wall Street Journal	Schools Want to Ban Phones. Parents Say No.	April 20, 2024	Education Department	reporters Thursday, top Education Department officials hailed the changes as the

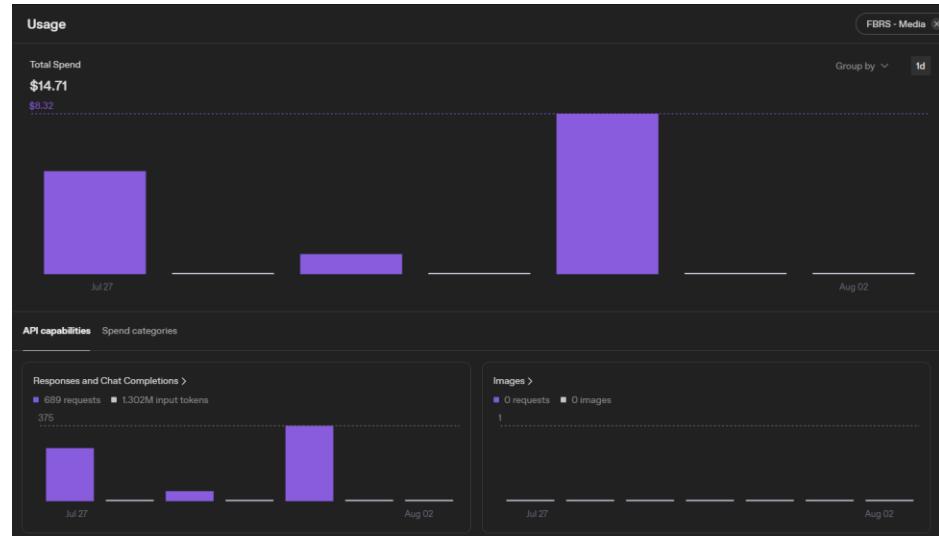
I tried probably every single OpenAI LLM available in July 2025 (pre-GPT-5, so that included 4o, o1, 4.1, 4.5, and o3) and most of them could pull out the metadata I wanted into a table, but only for the first 10 articles. The fact that it could do it, but not do it on 500 articles told me that it was possible to use AI, but we had to find a different way.

Step 2: Setting Up the Connection

Accessing OpenAI's models through its **API** **uses a different billing platform** than the chat interface

That means to use the API through a Bellwether account, you **have to contact the tech team to get set up**, which looks like:

1. Tech team created a new "Project" within the Bellwether workspace
 - this helps to categorize API calls, track usage, and allows bill projects accordingly
2. Developer (me) created an **API Key**
 - API Keys are used in every API request to tell OpenAI 1) who is sending the request, and 2) which project and workspace this belongs to



This is a sample report of my usage (only for the media analysis) from the week of 7/27/25 - 08/02/25). This type of disaggregated report is only possible if we get properly set-up on the backend.

 Each team member using the API should create their OWN API key and store it somewhere safe. **Do not share your API key, even among team members – that's like giving away your passport.**

Step 3: Writing the Prompt

The actual prompt sent to the GenAI model is in natural language.

- There might be certain parts of a prompt you might tweak (ex: respond in JSON format), but the message itself is written in plain English.

Write the prompt carefully.

- Because the prompt will likely stay the same for every API call, it becomes extremely important to have a well-crafted prompt that can account for edge cases
- For the media analysis, the initial prompts included input from the entire team, as well as refinements from ChatGPT itself

```
You are an expert research assistant. Extract the following metadata from the news article below.  
If a field is missing, fill it with N/A. Respond in JSON with these fields:  
- news_outlet  
- title  
- author(s) (comma-separated)  
- date_published  
- opinion_piece (news or opinion)  
- education_level (Preschool, K-12, postsecondary. if the article is about the Department of Education, tag as both K-12 and postsecondary)  
- states (states mentioned or covered, if any)  
- people_mentioned (names of individuals, comma-separated, with a one-word descriptor of their role in parentheses. For example, John Smith (teacher))  
- organizations_mentioned (names of organizations mentioned, comma-separated)  
- research_sources (names of orgs providing data, comma-separated. name the dataset in parentheses as well if applicable)  
  
Article:  
---
```

Step 4: Preparing the Request (coding)

Although the prompt uses natural language, it **requires code to transport the request through the API to the AI model.**

- **The language itself can vary:**
 - Most software developers would use languages like Java, C++, or Python.
 - Data analysts have our languages like R, Stata, or SPSS.
 - Some languages can span both use cases (Python).
- **But each language typically has a package of prebuilt codes that connect to APIs:**
 - For the media analysis, I used Python and its corresponding openai package
 - For R users, Alex has used the ellmr package



Unfortunately, APIs do not work without programming, so using code is not negotiable...

BUT



You can ask ChatGPT to write code for you, a process you might have heard referred to as "vibe-coding"



If you have *no experience coding*, you should get a second (human) opinion from someone familiar with the language, or you may inadvertently end up with security risks. But for those with some coding experience, using AI to debug or draft code is very useful and common.

Example: Python function to call API using prompt

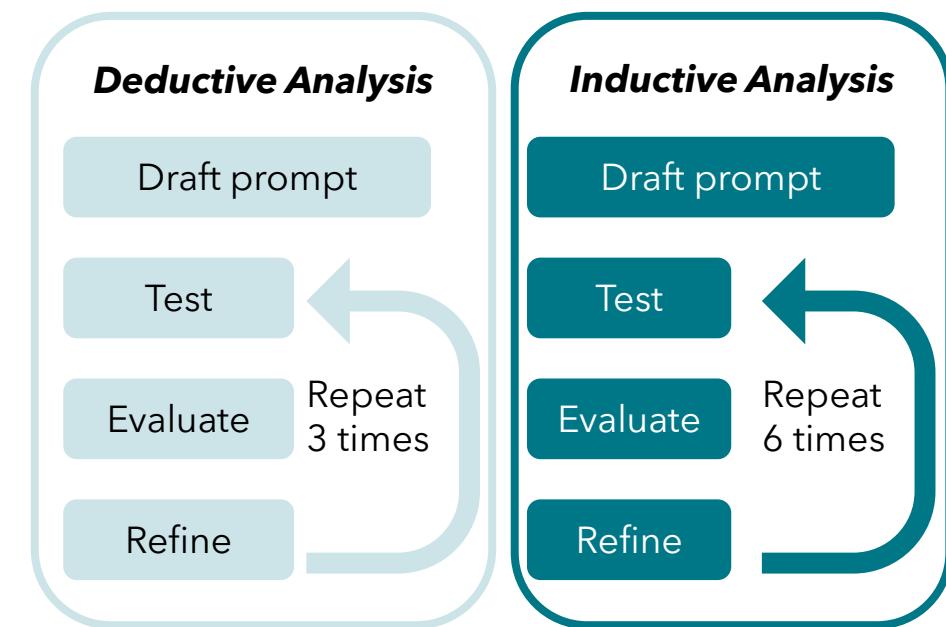
```
## Create function to call the API with the prompt
def extract_metadata_gpt41(article_text, filename):
    prompt = PROMPT_TEMPLATE.format(article_text=article_text[:75000]) # Limit 75k character input, change if needed
    content = ""
    try:
        response = client.responses.create(
            model=MODEL,
            instructions="You extract news article metadata.",
            input=prompt,
            max_output_tokens=400, # Limit 400 token output for cost, change if needed
            temperature=0
        )
        content = response.output_text.strip()
        print(f"Raw API output for {filename}:\n{content}\n")
        cleaned = clean_gpt_json(content)
        result = json.loads(cleaned)
        return result
    except Exception as e:
        print(f"Failed to process {filename}: {e}")
        print(f"Content was:\n{content}\n")
        print(f"Cleaned content:\n{cleaned}\n")
        return None
```

Step 5: Build and Conduct Testing Cycles

- 1. Choose a subset of your inputs to test:** this could be a random sample, or a carefully curated set that represents different use cases
- 2. Consider what you want from the AI outputs:**
 - Are the data straightforward enough that you can easily see if the AI is correct?
 - If not, what constitutes a “high-quality” output for each case?
- 3. Set a “minimum acceptable” threshold of quality:** How will you know if the output is “good enough”?
- 4. Run the test!**
- 5. Evaluate the outputs:** assess weaknesses, and refine your prompt or code as needed
- 6. Rinse and repeat #4-5 until you reach the desired threshold of quality**

Different types of outputs will require different levels of effort to both verify quality and refine prompting to get to “high quality.”

The media analysis had two workstreams with different outputs and different levels of testing:



If the analysis you’re conducting needs to be especially rigorous (e.g., for a field-facing publication), consult the evaluation team on what thresholds of quality are generally accepted.

Example: Evaluating Output Quality for Media Analysis

To the right are the results from our first prompt asking the AI to extract hard metadata (title, news outlet, region, etc.).

Green boxes are where AI got it correct, yellow boxes are where AI was semi-correct, and red boxes are where the AI was completely wrong.

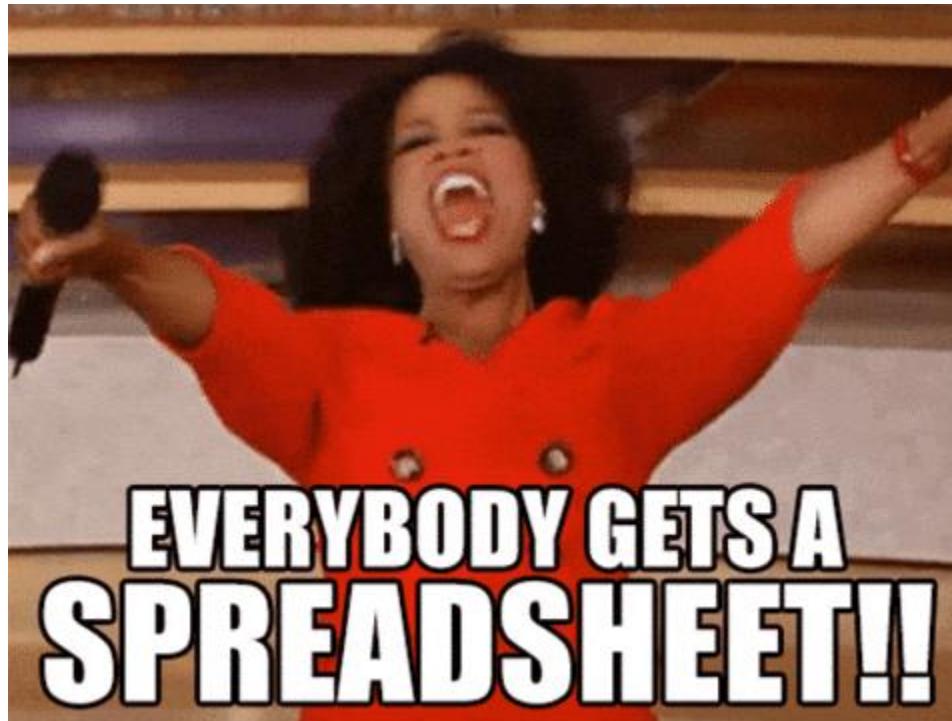
src_cnet	content	author	date_published	opinion_prc	education_level	status	data_sources	people_quoted
ABC News	Chad Guy Fieri celebrates birthday from his high I-94 with Kelly McCarty	6/9/2012	FALSE	K-12	[Education]	Approved	Stephanie Brad (abc), Lee Ann Wong (cbc), Curtis Luddo, Michael Symon (ch2), Jeff Lester (syfy), Brian Benben (nbc), Michael Symon (nbc)	
USA Today	Tonight's 'Dinner: 30' is here! See what's on the menu	6/9/2012	TRUE	K-12	[Entertainment]	Approved	Guy Fieri (dinner30), Brian Benben (nbc), Michael Symon (nbc)	
The New York Times	India Police Apprehend Suspect of Concealment of Body Parts	6/9/2012	TRUE	K-12	[Technology]	Approved	Prakash Singh (india), Michael Symon (nbc), Brian Benben (nbc)	
NPR	California school board rejects proposed notification of child's birth	7/6/2012	TRUE	K-12	[Culture]	Approved	David Greene (npr), Michael Symon (nbc), Brian Benben (nbc)	
Fox News	From modest debt to \$14 trillion, how Project 2025 could affect you	7/20/2012	TRUE	postsecondary	[California, Florida, New York]	Approved	Braddock Roberts (politicsnow), Joshua Zuckerman (advocacy), Bill Essey (essey), Michael Symon (nbc), Brian Benben (nbc)	
USA Today	U.S. says it may stop exports about abuses at kids' facilities in China	7/20/2012	TRUE	K-12	[Business]	Approved	Michael Symon (nbc), Brian Benben (nbc), Jeff Lester (syfy), Curtis Luddo (nbc), Michael Symon (ch2)	
USA Today	China says it may stop exports about abuses at kids' facilities in China	7/20/2012	TRUE	K-12	[Business]	Approved	Michael Symon (nbc), Brian Benben (nbc), Jeff Lester (syfy), Curtis Luddo (nbc), Michael Symon (ch2)	

To the left are the results after we refined the prompt. The amount of green has increased and yellow/red has decreased, signaling that the new prompt leads to higher quality outputs.

Step 6: Full Send 🥰

For the media analysis, we had a quality threshold of 80% accuracy, as compared to human answers, which is based on a standard accepted widely in social sciences for content analyses (shout-out Nora!).

Once we hit that threshold, we ran both prompts on the full set of 1579 articles, and in return got lovely .csv files, which Nora then analyzed further to distill trends.

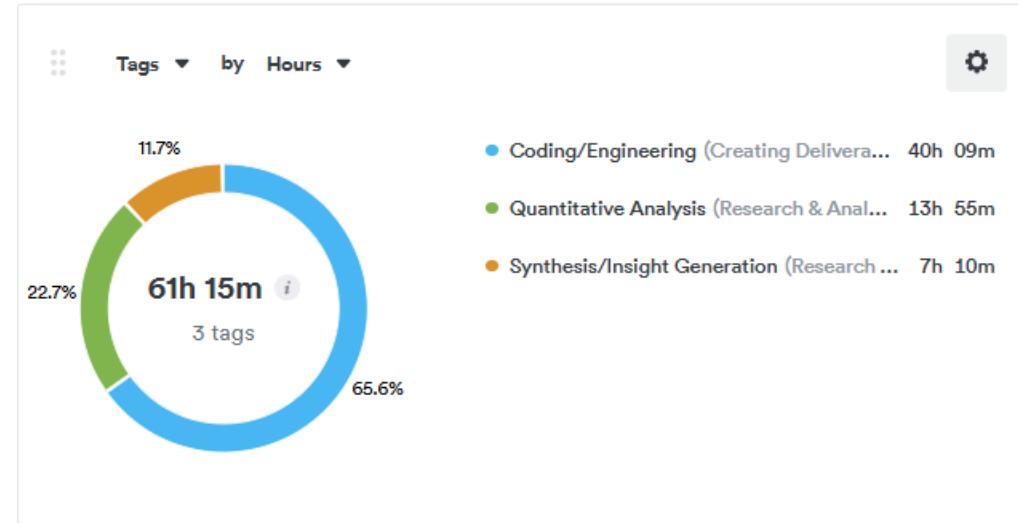


Seems like a lot of work. Is it worth it?

I spent ~40 hours coding the API calls, and ~60 hours total on drafting, testing, evaluating, and refining the analysis + its outputs.

Nora likely spent another ~20 helping with evaluating outputs and refining the prompts.

Total Hours Clocked with AI: 80



Without AI, however, Nora and I would have been reading each article manually and extracting metadata and themes into a spreadsheet.

We did this for a sample of articles, and I averaged ~10 mins/article.

Total Hours Likely Required without AI: 263

~10 mins. per article

X 1,579 articles

15,790 minutes

/ 60 mins. per hour

263 hours

Q&A

