



# ScrapeGraphAI

You Only Scrape Once

Software Developer's Thursday Edition

**NOI TECHPARK BOZEN-BOLZANO**



**NOISE**



**1ST AUGUST 2024**

# Our Team



**Marco Vinciguerra**

MSc Computer Engineering

–

**Co-founder @  
ScrapeGraphAI, Inc**



**Marco Perini**

MSc Mechatronics Engineering

–

**Co-founder @  
ScrapeGraphAI, Inc**

–

**Researcher @  
Eurac Research**



**Lorenzo Padoan**

MSc Computer Science

–

**Co-founder @  
ScrapeGraphAI, Inc**

# The Era of Big Data

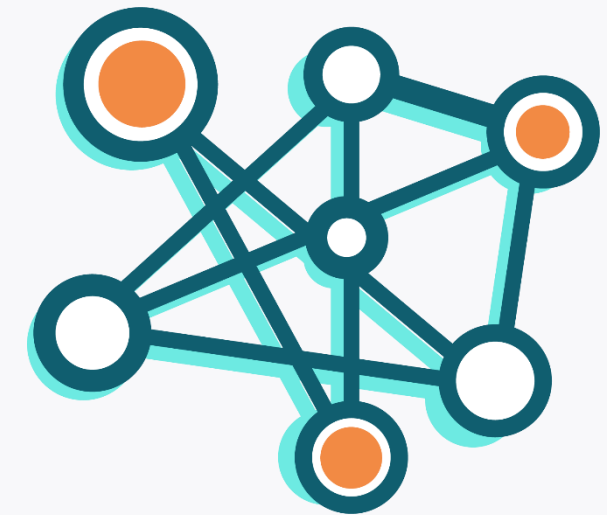


**Analytics**

Gemini



**LLM Training**



**Data Connections**

# Internet



Principal source of data

# What is a Scraper



**Scraping** is the act of **extracting** information  
from a **data source**



# Common scraping tools



Dev  
tools

Beautifulsoup



Web  
services

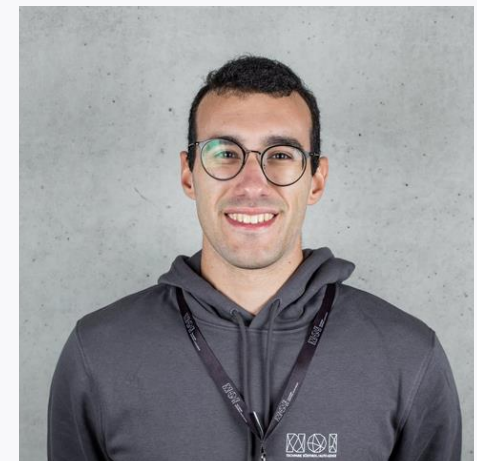
 Web  
Scraper

  
Octoparse

# Our Question



Is it possible to **scrape** websites **without** any knowledge of **HTML**, just by **writing what I want and how we want it?**



# Our Solution



Yes 🎉 🎉 🎉  
with



ScrapeGraphAI



# Our Solution



## ScrapeGraphAI



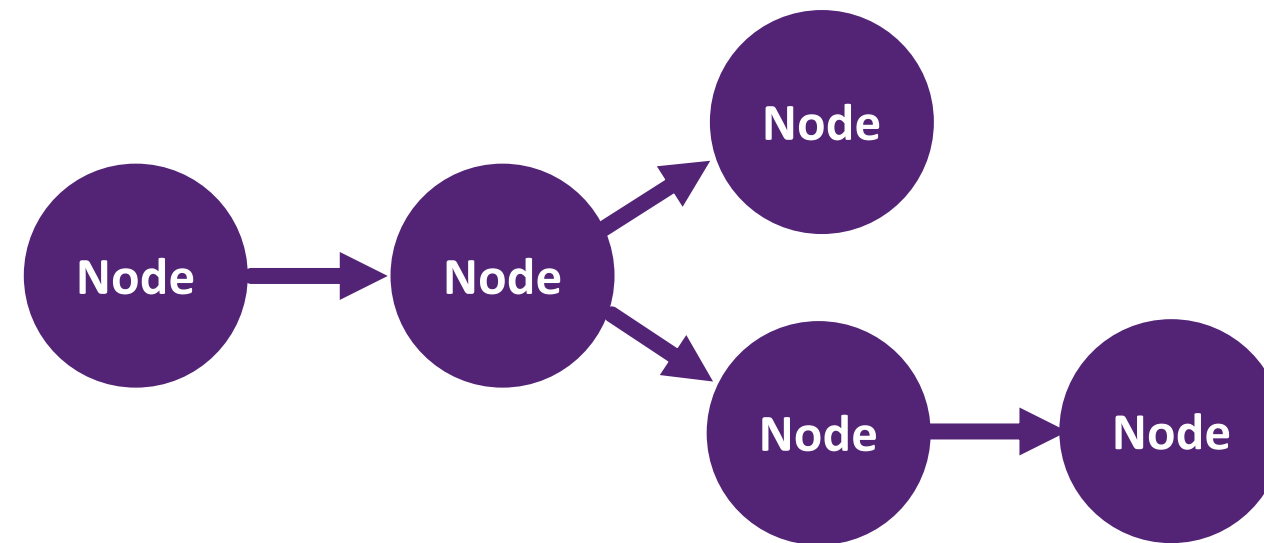
# Why SGAI?



**Scrape** →

Extract information from the **web**

**Graph** →



**AI** →

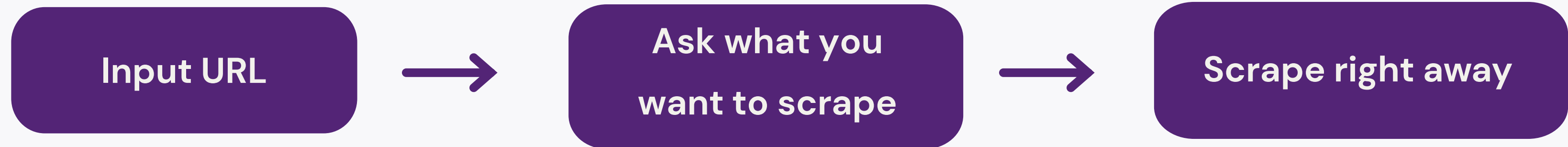
Use LLMs to analyze the **website structure and content**

# Traditional Workflow



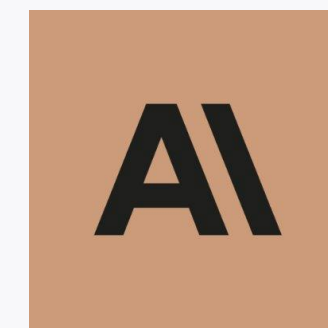
- It doesn't adapt to website **structure changes**
- Custom code for **different websites**

# SGAI Workflow

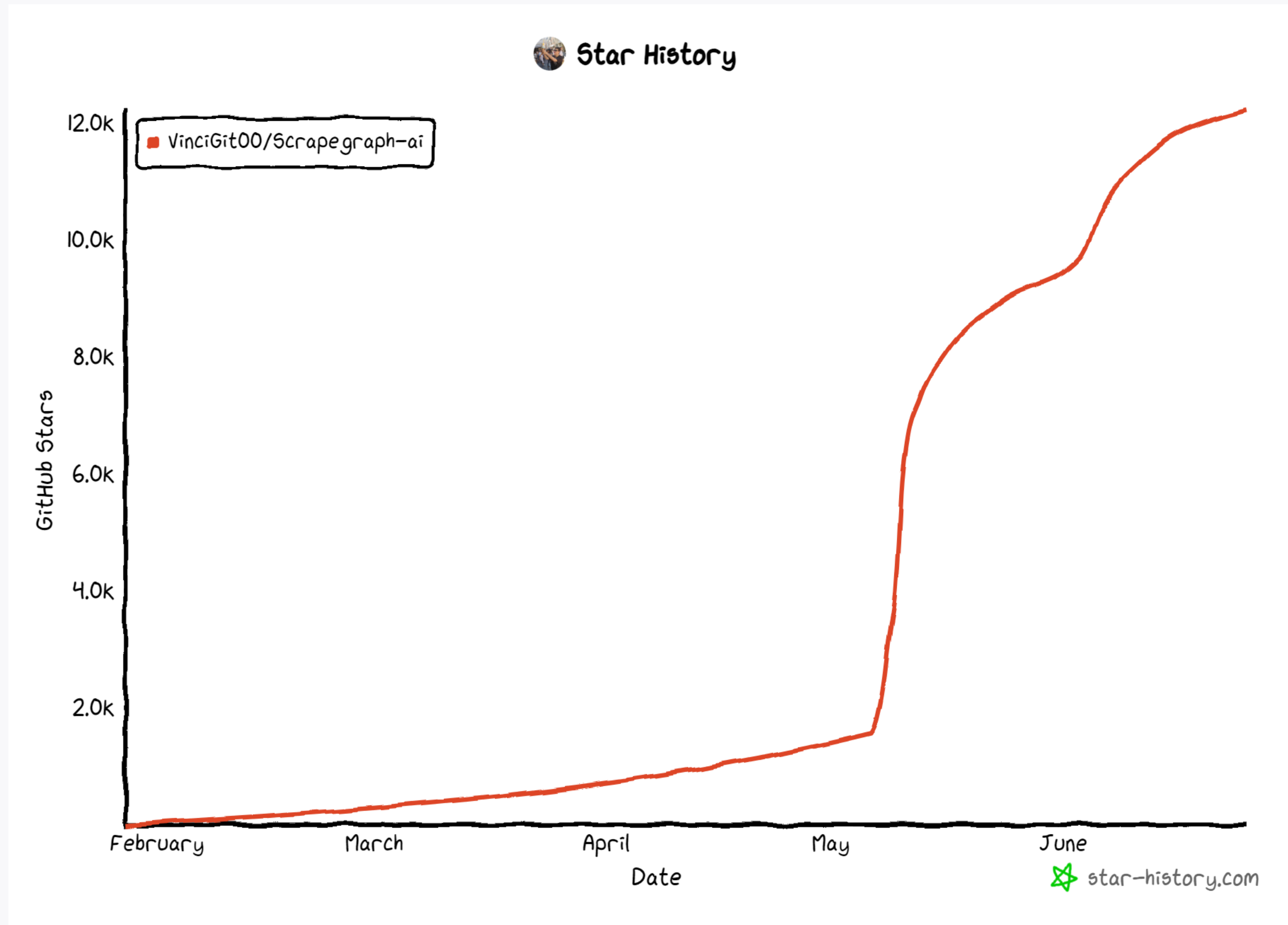


- Adapts to website **structure changes**
- **Corrects itself** until it succeed
- Flexibility in scraping **different websites**

# Available LLMs

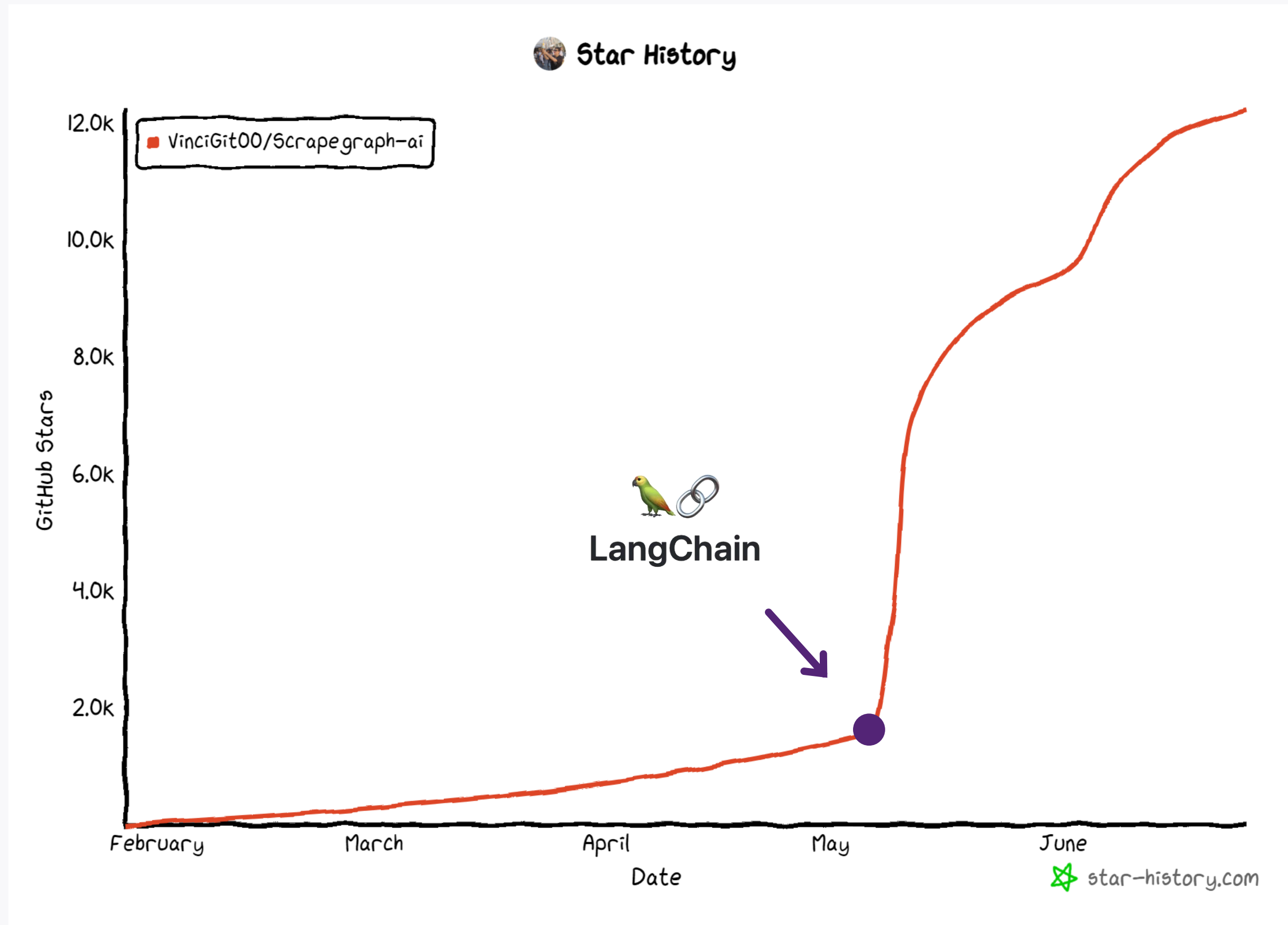


# Growth





# Growth






# Acceleration



- Python library + documentation
- LinkedIn + Twitter for updates
- Discord for developers and support
- Participated to Hackathons
- Using trending technologies
- Actively maintaining the project



**LangChain**  
243,516 followers  
3mo • 

 ScrapeGraphAI: You Only Scrape Once

ScrapeGraphAI is a web scraping python library that uses LLMs to create scraping pipelines for websites, documents and XML files. Just say which information you want to extract and the library will do it for you!

<https://lnkd.in/gDKxS62h>

**Case 1: Extracting information using Ollama**




Remember to download the model on Ollama separately!

```
from scrapegraphai.graphs import SmartScraperGraph

graph_config = {
  "llm": {
    "model": "ollama/mistral",
    "temperature": 0,
    "format": "json", # Ollama needs the format to be specified explicitly
    "base_url": "http://localhost:11434", # set Ollama URL
  },
  "embeddings": {
    "model": "ollama/nomic-embed-text",
    "base_url": "http://localhost:11434", # set Ollama URL
  }
}

smart_scraper_graph = SmartScraperGraph(
  prompt="List me all the articles",
  # also accepts a string with the already downloaded HTML code
  source="https://perinim.github.io/projects",
  config=graph_config
)

result = smart_scraper_graph.run()
print(result)
```

   You and 3,463 others

133 comments • 269 reposts

# Featured by



# Open-Source Community



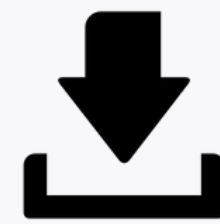
After 4 months of development:



**13k+**  
stars on Github



**1k+**  
forks

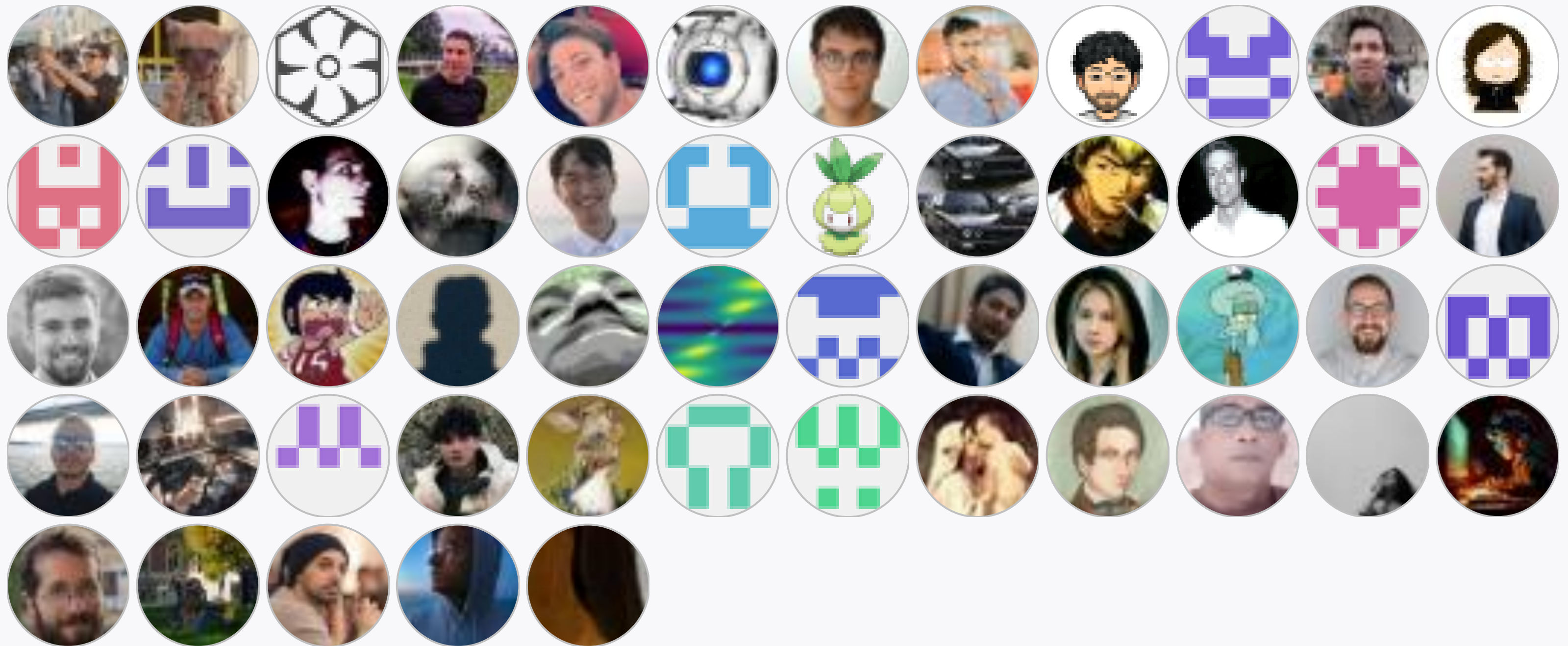


**150k+**  
downloads  
on pypi (pip)



**500+**  
users  
on Discord

# Contributors





# Technical Specifications





Source

`https://example.com/`

Prompt

List me what's inside the page, URLs and provide a summary



ScrapeGraphAI



Extracted Data

Content: ...  
URLs: [...]  
Summary: ...



### Available Scraping Pipelines

SmartScraper

Speech

MDScraper

CSVScraper

OmniScraper

Search

ScriptCreatorMulti

PDFScraper

JSONScraper

OmniSearch

# Nodes (node)



## Fetch

Fetch **HTML** from **URL** or **local file**

## Parse

Parse input **document** and **split** it in **chunks**

## RAG

Chunks **vector storage** and **retrieval**

## Generate Answer

**Analyze** chunks and **merge** the **results**

## Robots

Analyze **robots.txt** and check **scrapability**

## Search Internet

Get **search** engine **results** from **prompt**

## Search Links

Get **URLs** from a **webpage**

## TextTo Speech

**Convert** input **text** to **audio**

## ImageTo Text

**Convert** **image** to **text** from **URL**

# Graphs ( node → node )



**SmartScraperGraph**



One page scraper

**SearchGraph**



Multi-page search engine scraper

**ScriptCreatorGraph**



Python script generator

**OmniScraperGraph**



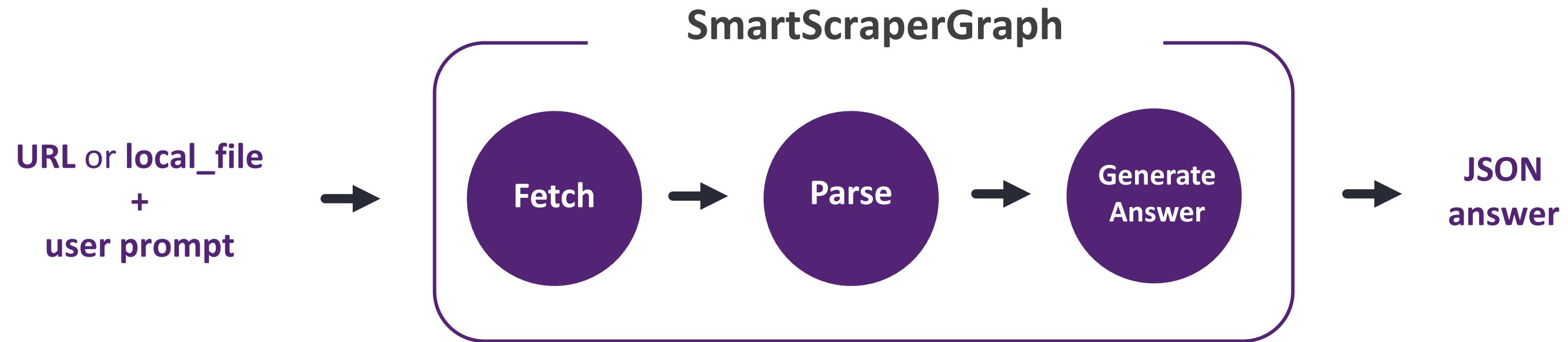
One-page Multi-modal scraper

**SpeechGraph**

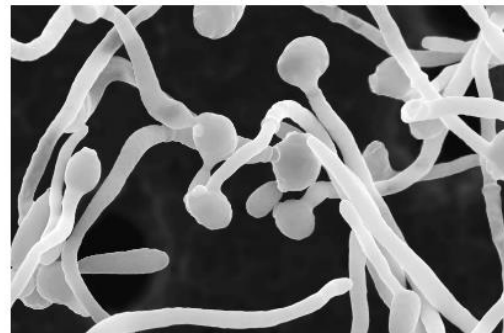


Generate audio answer

# SmartScraperGraph



## HEALTH



### Unruly Gut Fungi Can Make Your Covid Worse

An infection can upset your microbiome, and if certain gut fungi run riot, this can kick the immune system into overdrive.

MAGGIE CHEN



### Bird Flu Is Spreading in Alarming New Ways

H5N1 has infected cattle across the US and jumped from a mammal to a human for the first time. Experts fear it may someday evolve to spread among humans.

DAVID COX



### This Woman Will Decide Which Babies Are Born

Noor Siddiqui founded Orchid so people could “have healthy babies.” Now she’s using the company’s gene technology on herself—and talking about it for the first time.

JASON KEHE

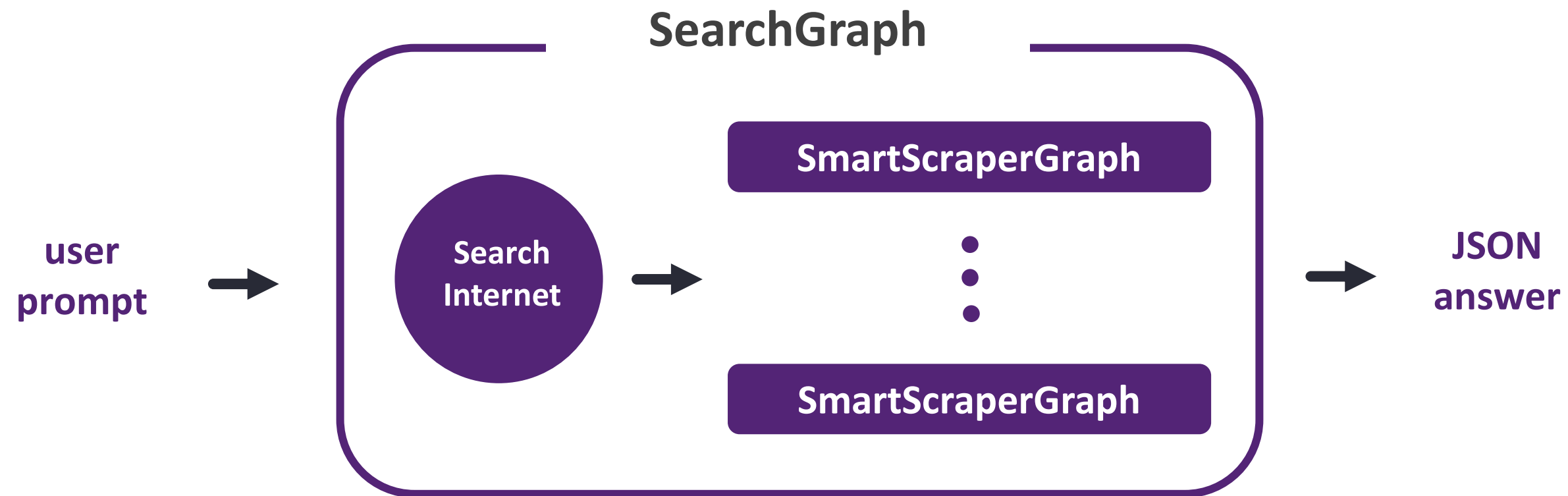


### ‘In 24 Hours, You’ll Have Your Pills’: American Women Are Traveling to Mexico for Abortions

Since the US Supreme Court overturned *Roe v. Wade* in 2022, more women have been crossing the border to Mexico for abortion medications and procedures.

CARMEN VALERIA ESCOBAR

# SearchGraph



Sottomarina.net

<https://www.sottomarina.net> › g... · Traduci questa pagina

## Culinary traditions and cuisine of Chioggia-Sottomarina

**MAIN COURSES.** • Bisato in tecia: eel in tomato sauce and white wine. • Sepe nere: cuttlefish, boiled in a mixture of onion and garlic, with the addition of ...



Museo di Zoologia Adriatica Giuseppe Olivi

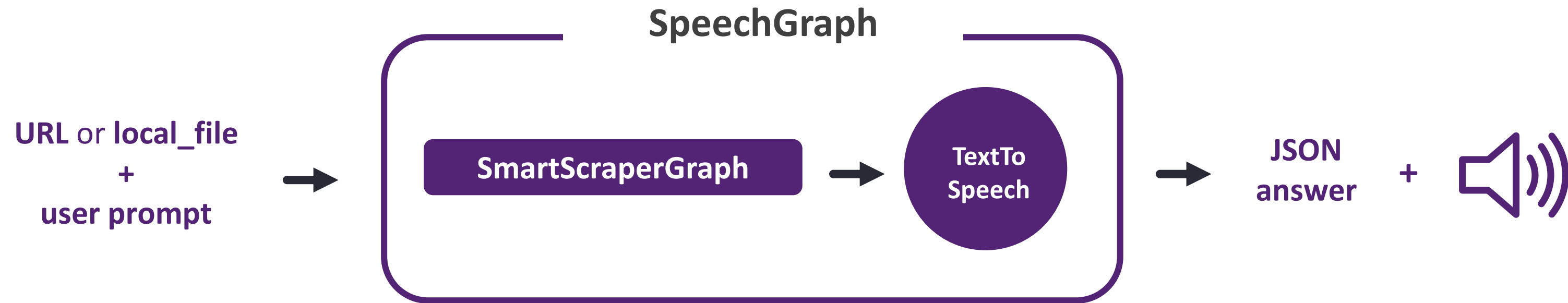
<https://www.museoolivi.it> › typi... · Traduci questa pagina

## Typical Food and Delicacies | Museo G. Olivi Chioggia

The most famous and renowned vegetable cultivated in the area is the red radicchio of **Chioggia** (*Cichorium intybus*), also called "Rose of **Chioggia**", exclusive ...



# SpeechGraph



## SARDE IN SAOR: RICETTA ORIGINALE VENEZIANA

*Fanno parte della cucina veneta ma sono conosciute ed amate anche in altre regioni d'Italia. Le sarde in saor, infatti, costituiranno una piacevole scoperta per chi poco conosce la tradizione veneziana. Ecco la ricetta originale e tanti consigli utili*

Difficoltà: Preparazione: 25 min Cottura: 10 min Porzioni: 4 persone Costo:



# Pydantic Schema



```
class News(BaseModel):
    title: str = Field(description="The title of the news")
    description: str = Field(description="The description of the news")

class ListNews(BaseModel):
    news: List[News] = Field(description="List of news")

result = SmartScraperGraph(
    prompt="List me all the AI related news with their description.",
    source="https://www.wired.com/category/science/",
    config=graph_config,
    schema=ListNews,
).run()
```

```
{
  "news": [
    {
      "title": "Light-Based Chips ...",
      "description": "Optical neural networks ..."
    },
    ...
    {
      "title": "AI Is Building ...",
      "description": "Robots, computers, and ..."
    }
  ]
}
```

# Comparative Results



Let's suppose we want to extract the news titles from  
<https://www.wired.com>

The screenshot shows the Wired.com homepage. At the top, there's a navigation bar with the Wired logo and various category links: SECURITY, POLITICS, GEAR, BACKCHANNEL, BUSINESS, SCIENCE, CULTURE, IDEAS, and MERCH. There are also links for SIGN IN and SUBSCRIBE. The main content area features a large banner for a Leica camera with the text "M è Mentalità." and a red button that says "Scopri di più". Below the banner, there are two columns of news articles. The left column is titled "TODAY'S PICKS" and features an article about LinkedIn with the headline "LinkedIn Tells People if You Look at Their Profile. Here's How to Turn That Off" by Justin Pot. The right column is titled "MOST RECENT" and features three articles: "I Found Frank Herbert's Dune Script. Dune: Part Two Is Better" by Max Evry, "JavaScript Runs the World—Maybe Even Literally" by Sheon Han, and "Less Sea Ice Means More Arctic Trees—Which Means Trouble" by Matt Simon. At the bottom of the right column, there's a partial view of an article about Robert F. Kennedy Jr.'s Microsoft-powered project.

# Comparative Results



## BeautifulSoup

```
import requests
from bs4 import BeautifulSoup

# Richiesta alla pagina web
url = "https://www.wired.com/"
response = requests.get(url)

# Controllo del codice di stato
if response.status_code == 200:
    # Parsing del contenuto HTML
    soup = BeautifulSoup(response.content, "html.parser")

    # Trova tutti i titoli degli articoli
    articoli = soup.find_all("h2", class_="title")

    # Estrai il testo di ogni titolo e autore
    nomi_articoli = [articolo.text for articolo in articoli]
    autori = []

    # Trova gli autori per ogni articolo
    for articolo in articoli:
        autore_el = articolo.find_next_sibling("span", class_="byline-author")
        if autore_el:
            autori.append(autore_el.text)
        else:
            autori.append("Autore non trovato")

    # Stampa i nomi degli articoli e gli autori
    for nome, autore in zip(nomi_articoli, autori):
        print(f"Articolo: {nome}")
        print(f"Autore: {autore}")
        print()
else:
    print("Errore durante la richiesta:", response.status_code)
```

## ScrapeGraphAI

```
from scrapegraphai.graphs import SmartScraperGraph
OPENAI_API_KEY = "YOUR_API_KEY"

# Define the configuration for the graph
graph_config = {
    "llm": {
        "api_key": OPENAI_API_KEY,
        "model": "gpt-3.5-turbo",
    },
}

# Create the SmartScraperGraph instance
smart_scraper_graph = SmartScraperGraph(
    prompt="List me all the news with their description.",
    file_source="https://perinim.github.io/projects/", # also accepts a string with the all
    config=graph_config
)

result = smart_scraper_graph.run()
print(result)
```

<concise /> <less code /> <reusable />

For scraping another website, you just need to change **2 lines!!**

# Pros of ScrapeGraphAI



- **Low code** and **fast** implementation
- **Fault tolerant** to dynamic HTML code
- Possibility to run **local LLMs**
- **Portability**
- **Open-Source community**



# Benchmarks



**Prompt:** List me all the news with their descriptions

Model	ansa.it	wired.com
Macbook 14' m1 pro (Mistral on Ollama with nomic-embed-text)	16.291s	38.74s
gpt-3.5-turbo	4.132s	8.836s
gpt-4-turbo-preview	6.965s	21.53s
gpt-4o	4.446s	15.27s
Groq	1.335s	5.32s

# Integrations



🏠 / project / smart-scraper-example / unique-id-0

<	Action	Ran	Duration	+ Spans	Live	Data	Action
0	Fetch	5/27/2024, 7:56:15 AM			running	State -	

```
graph TD; Fetch[Fetch] --> Parse[Parse]; Parse --> RAG[RAG]; RAG --> GenerateAnswer[GenerateAnswer];
```

React Flow



# Repositories



**ScrapeGraphAI**

+



**Documentation**

`ScrapeGraphAI/Scrapegraph-ai`

`docs.scrapegraphai.com`

+

`scrapegraph-ai.readthedocs.io`

If you like the project, feel free to **leave a star**



# Library Roadmap



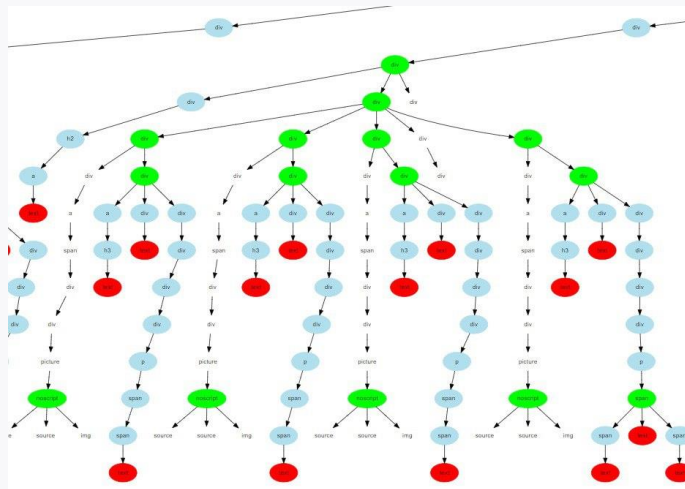
DeepSearch  
Graph

Screenshot  
Scraping

New  
Webdrivers

Page  
Caching

Captcha  
Solving



# SGAI Companion



vinci00

```
/llama-3-8b-Instruct-bnb-4bit-  
scrapegraph-companion
```

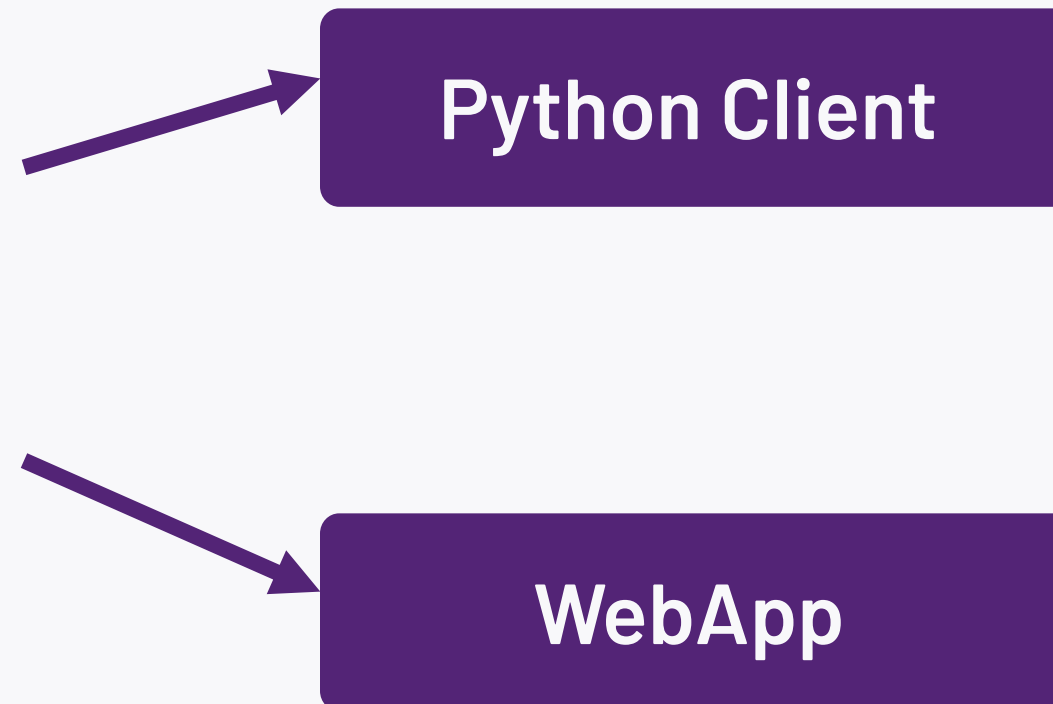


huggingface.co

# API Rollout



<https://api.scrapegraphai.com/>



Looking for  
**Early Adopters**  
(scan the QR-Code)

# Thank you for the attention!



## ScrapeGraphAI

You Only Scrape Once

[contact@scrapegraphai.com](mailto:contact@scrapegraphai.com)