

# HOW TO DO DATA ANALYSES PROFESSIONALLY

**ChatGPT or Claude makes you think  
you are doing ~~DATA ANALYSES~~  
but you're NOT**

# What will you learn

- Why AI Chat Isn't Real Data Analysis
- AI Data Processing Fundamentals
- Common Failure Patterns and Solutions
- Reliable Data Processing Workflow with Chatbots
- Big Datasets: Scaling beyond chat bots
- Model Selection and Cost Optimization
- Advanced Prompt Engineering for Data Analysis
- Professional Dashboard Creation

# **Skills & tools you learn and use**

## **Skills**

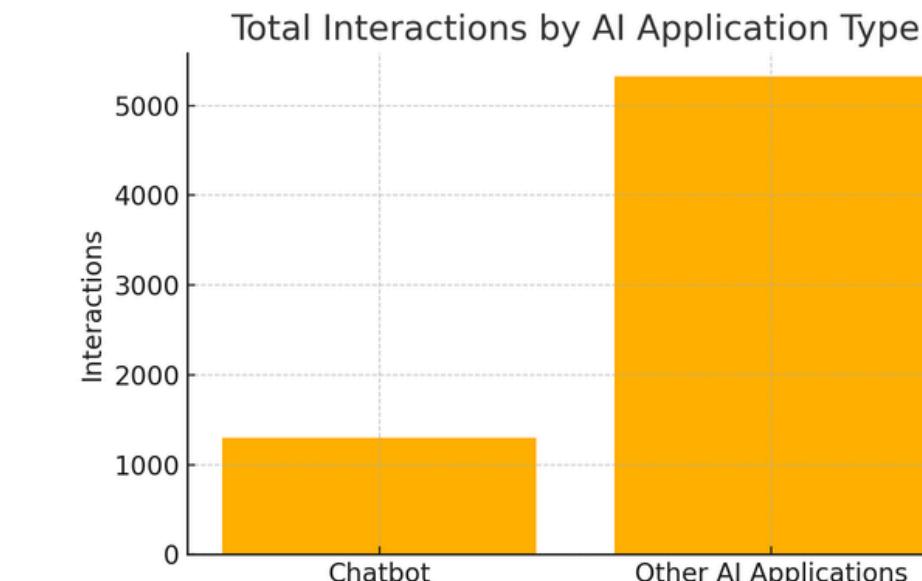
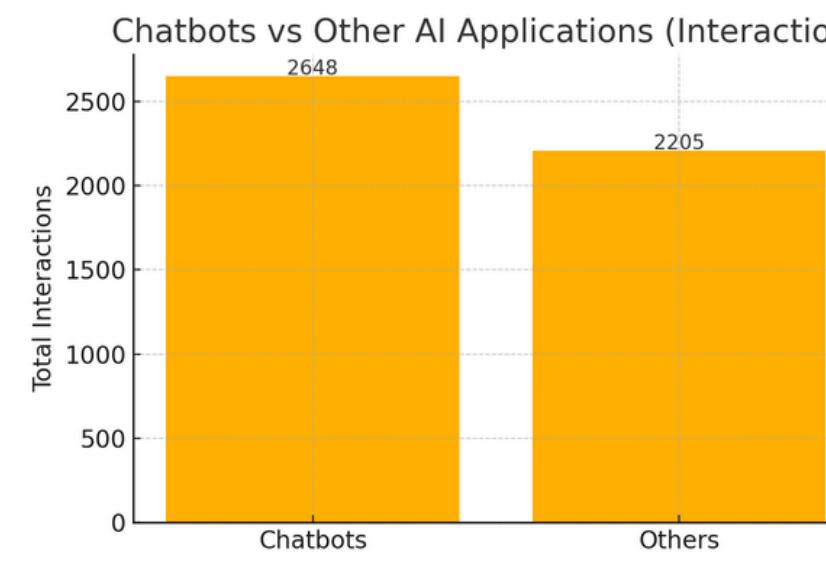
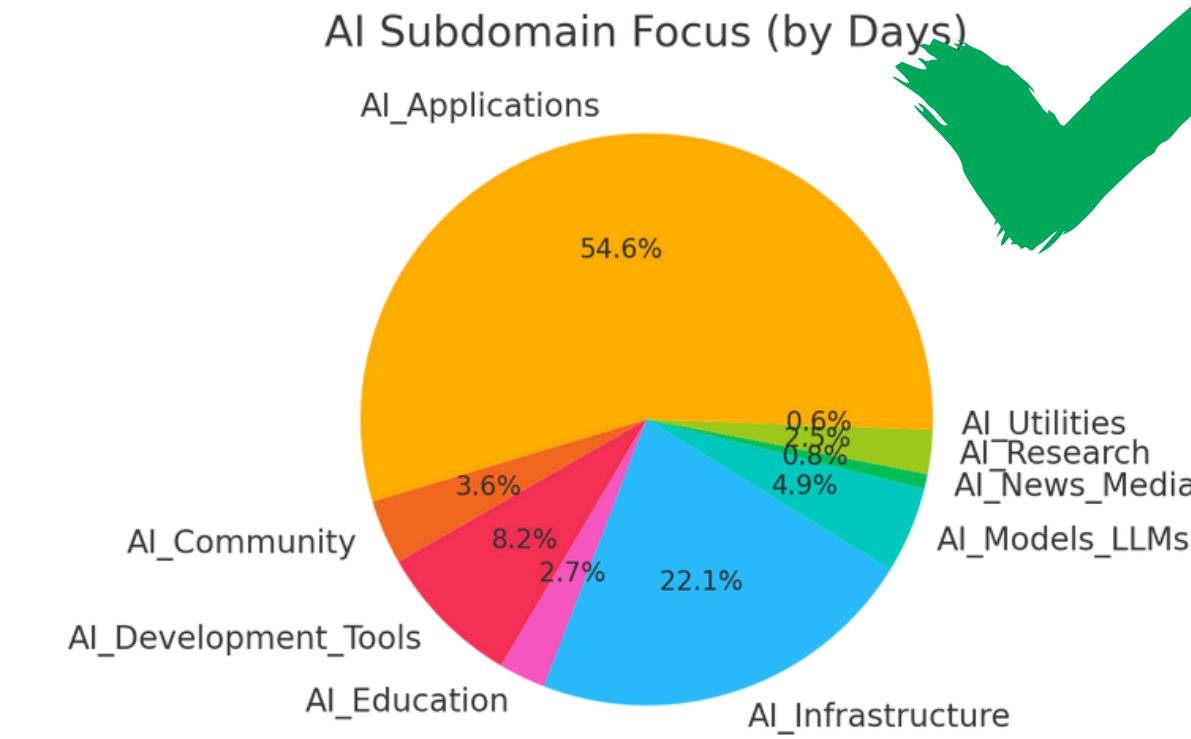
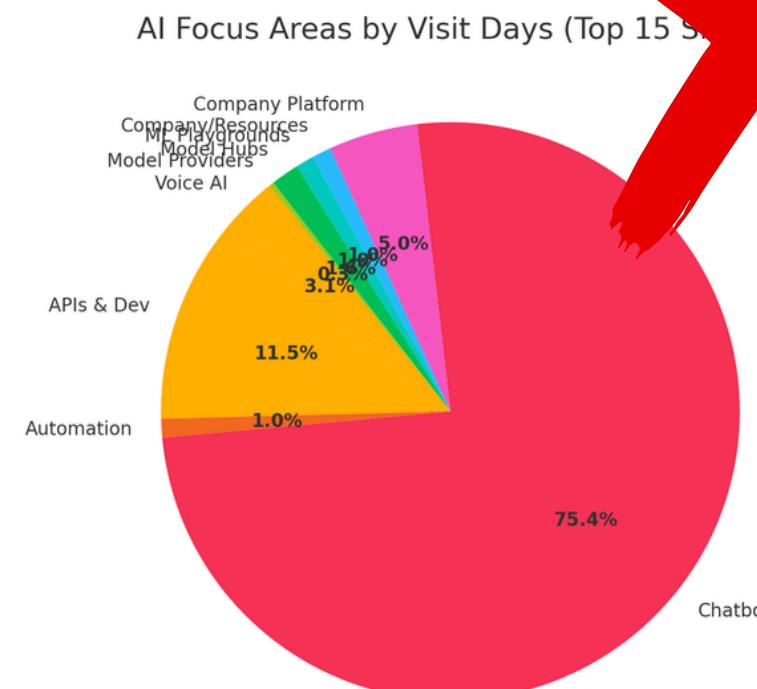
- 1. Technical Implementation: Custom functions, API integration**
- 2. Data Processing: Data cleaning, enrichment, and validation workflows**
- 3. Model Management: Selection, evaluation, and optimization across multiple LLMs**

## **Tools & Technologies**

- 1. OpenRouter API platform with 200+ AI models**
- 2. Excel VBA Script integration**
- 3. Professional prompt engineering techniques**
- 4. Cost optimization**
- 5. Data validation and edge cases**

# This may happen to you, and you don't even know it

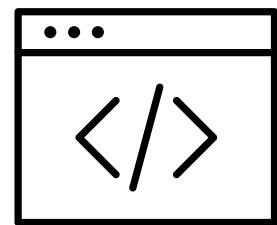
## Same Dataset. Different results



# **AI Data Processing Fundamentals**

# Data Analytics with AI: Methods

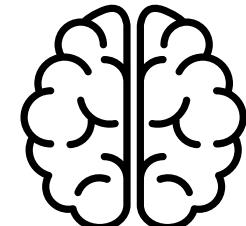
programmatically



**Processing data faster**

**Get reports and charts faster**

direct generation



**Enriching the data**

**Getting new insights**

# Dataset size

vs.

# Dataset content

vs.

# Analytics use case

DATA SCALE	VOLUME	STRATEGY	SUCCESS FACTOR
Small	< 100 records	Direct Prompting	Single-pass analysis
Medium Structured	100s-1000s records	Code Generation	Programmatic edge case handling
Medium Unstructured	100s-1000s documents	Iterative API Processing	Document-by-document consistency
Medium Hybrid	100s-1000s mixed data	Two-Step Pipeline	Enrichment → Analysis separation
Large Scale	10K+ records	Batch API Processing	Chunked processing with error handling
Large Corpus	1000s+ documents	Map-Reduce Approach	Parallel processing → Synthesis

# **Common Failure Patterns and Solutions**

# Common Problems



## Programmatic Approach

### Code Hallucination

AI generates syntactically correct code that's logically wrong. Assumes data structure without actually inspecting it. Creates hard-coded values instead of dynamic discovery.

### Sample Bias Issues

AI only sees the first few rows you show it. Misses rare categories, edge cases, or data quality issues. Generates code that breaks on unseen data variations.



## Direct Generation

### Context Window Ceiling

Can't fit 1000+ records into a single prompt. Forces you to chunk data, losing global context. Analysis becomes fragmented and inconsistent.

### Output Truncation

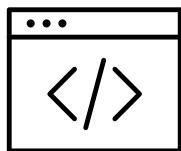
Responses get cut off mid-analysis. Complex charts/reports become incomplete. Lost insights at the end of processing.

### Consistency Breakdown

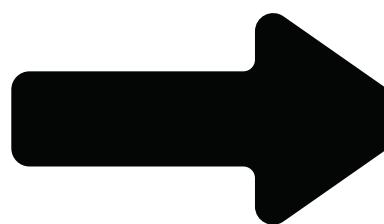
Different analysis quality across data chunks. Can't maintain analytical thread across multiple API calls. Contradictory insights from different processing sessions.

# AI role in Data Analytics

programmatically

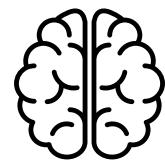


**Processing data faster**  
**Get reports and charts faster**

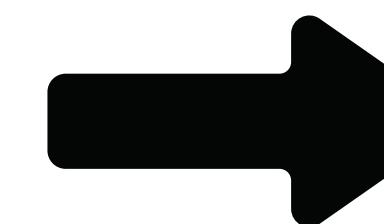


**You could do with Excel  
given enough time and skill**

direct generation

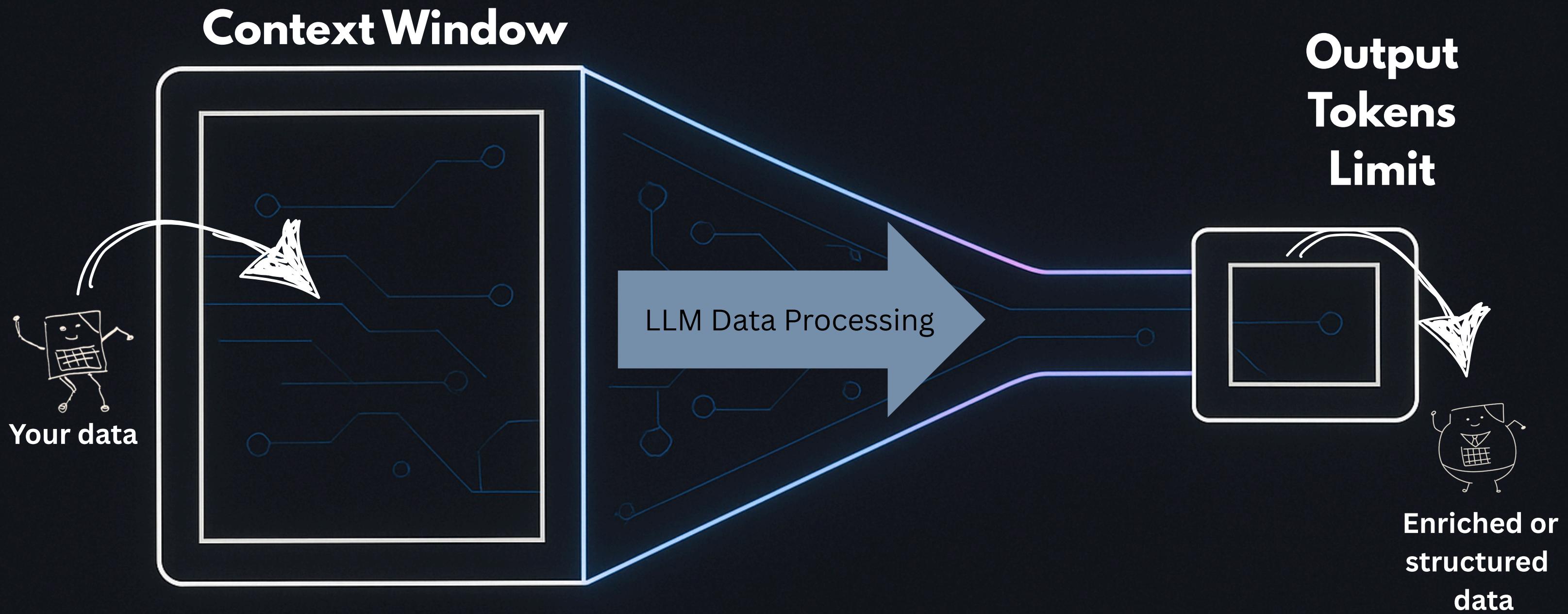


**Enriching the data**  
**Getting new insights**

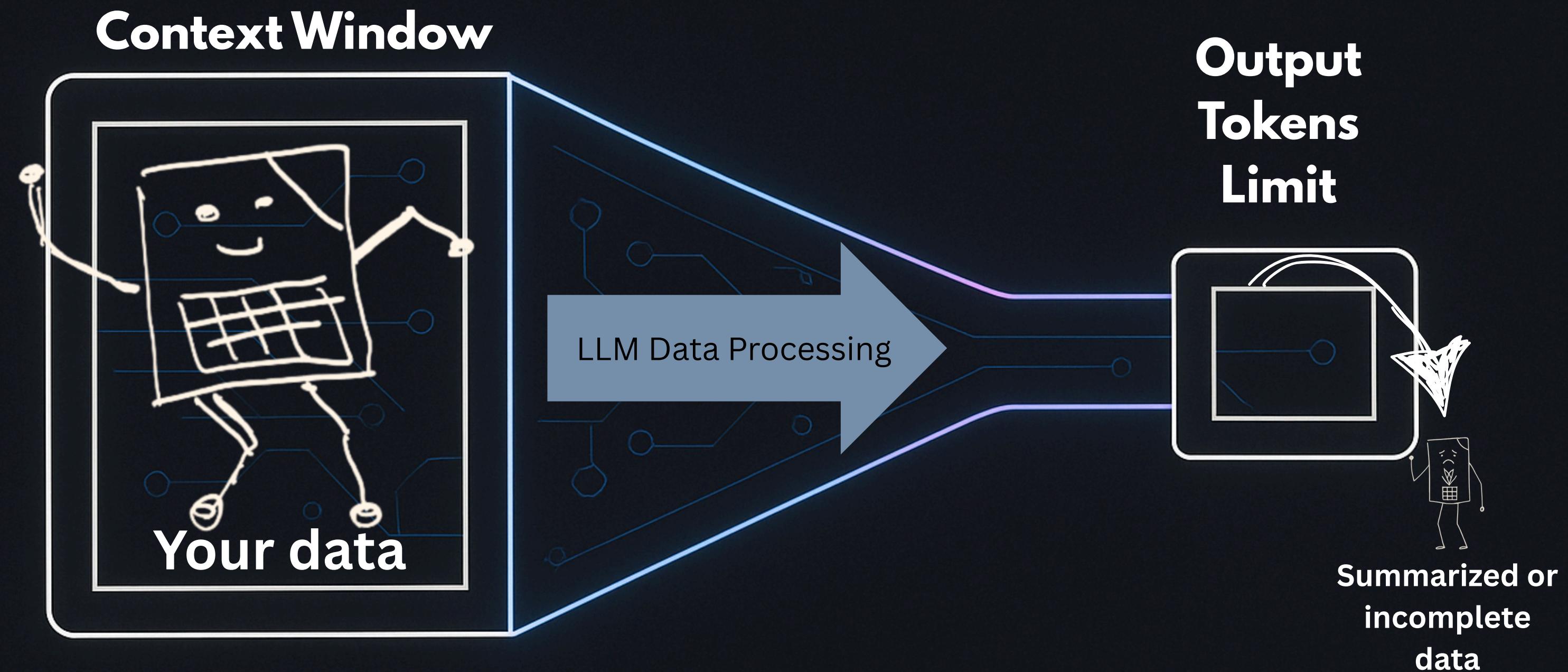


**This is specific to AI  
You cannot do this other way**

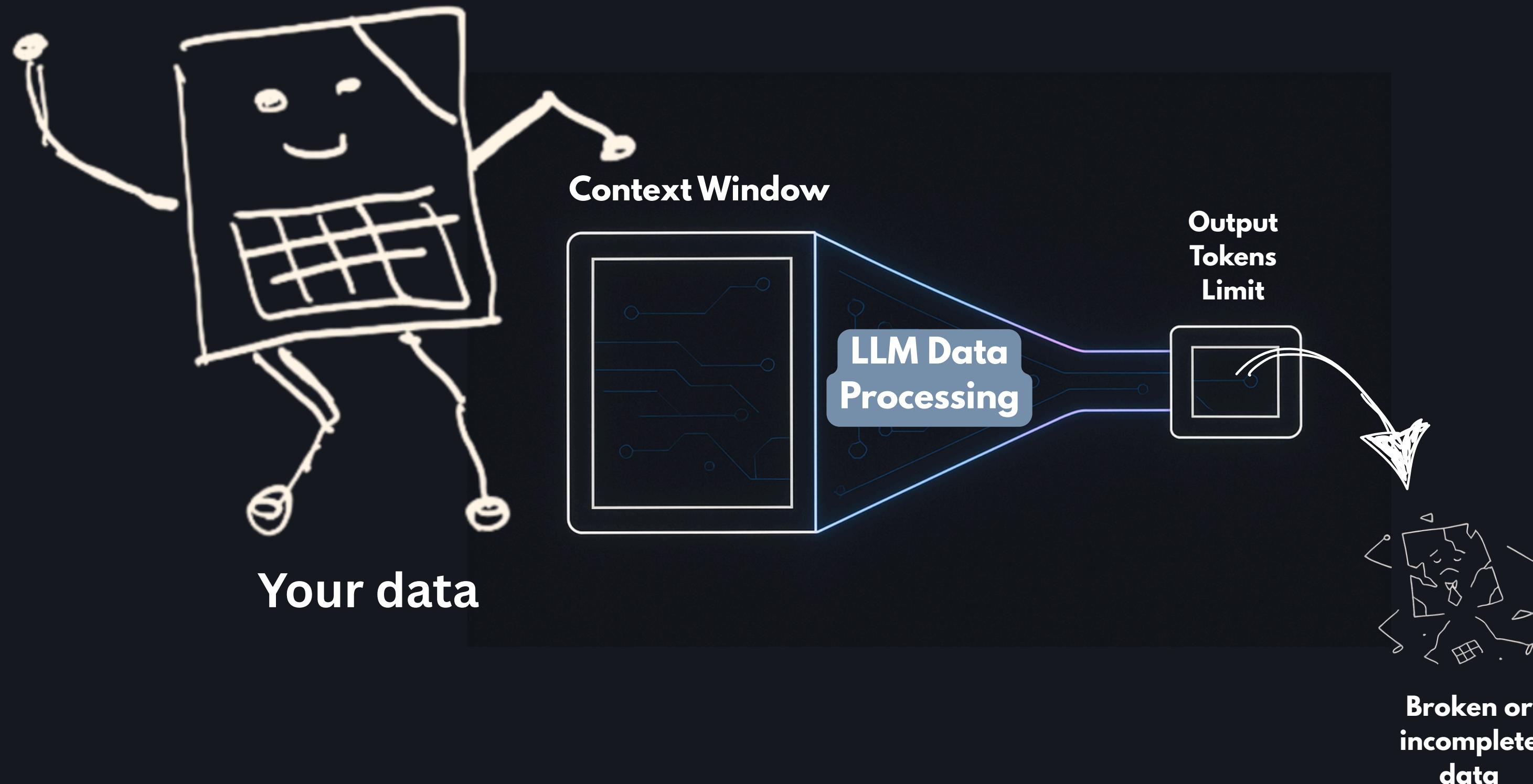
# Unstructured data processing



# Unstructured data processing



# Unstructured data processing

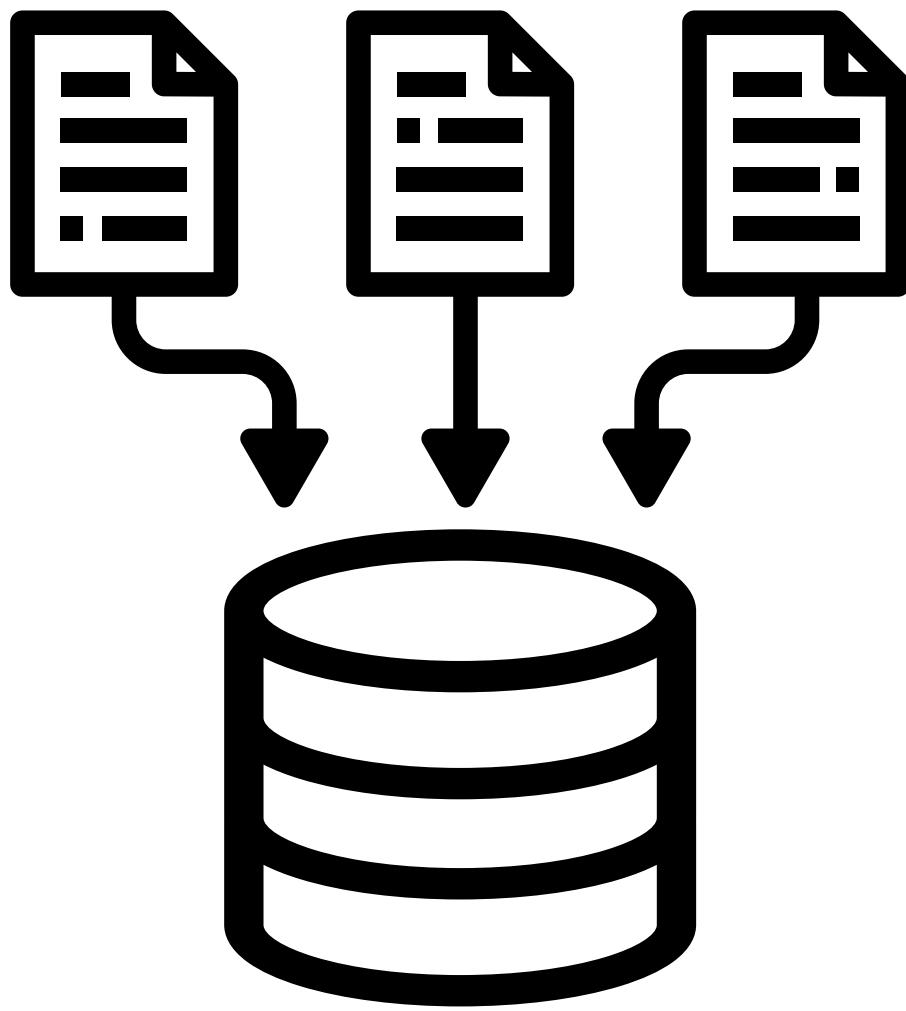


# AI Model Comparison

Context Windows, Output Limits & Performance Ratios

Provider	Model	API Context	API Output	Chat Context	Chat Output	API I/O Ratio
OpenAI	GPT-4o	128,000	16,384	32,000	≈4,096	7.8
	GPT-4o-mini	128,000	16,384	32,000	≈4,096	7.8
	GPT-4.1	1,047,576	32,768	32,000	≈4,096	32.0
	GPT-4.1 mini / nano	1,047,576	32,768	32,000	≈4,096	32.0
	GPT-4.5 <small>preview</small>	128,000	≈8,192*	32,000*	≈4,096*	≈15.6
	o3 (reasoning)	200,000	100,000	—	—	2.0
	o4-mini	200,000	100,000	≈32,000*	≈4,096*	2.0
Google	Gemini 2.5 Pro	1,048,576	65,536	1,000,000 (Advanced) / 32,000 (Free)	≈8,192*	16.0
	Gemini 2.5 Flash	1,048,576	8,192	1,000,000 (Advanced) / 32,000 (Free)	≈8,192*	128.0
Anthropic	Claude 4 Sonnet	200,000	64,000	200,000	64,000	3.1
	Claude 4 Opus	200,000	32,000	200,000	32,000	6.3
Mistral	Mistral Medium 3	128,000	8,192*	128,000*	8,192*	15.6
DeepSeek	DeepSeek-R1 (reasoner)	64,000	≤64,000	4,000 (free chat)	≈4,000	1.0
Meta	Llama 4 Scout	10,000,000	— (not published)	—	—	—

For premium account some limits might be higher in the chat

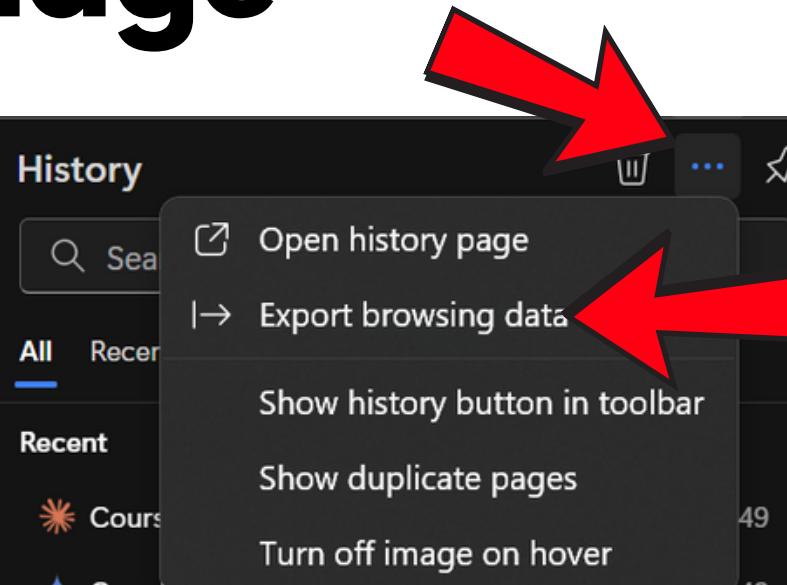
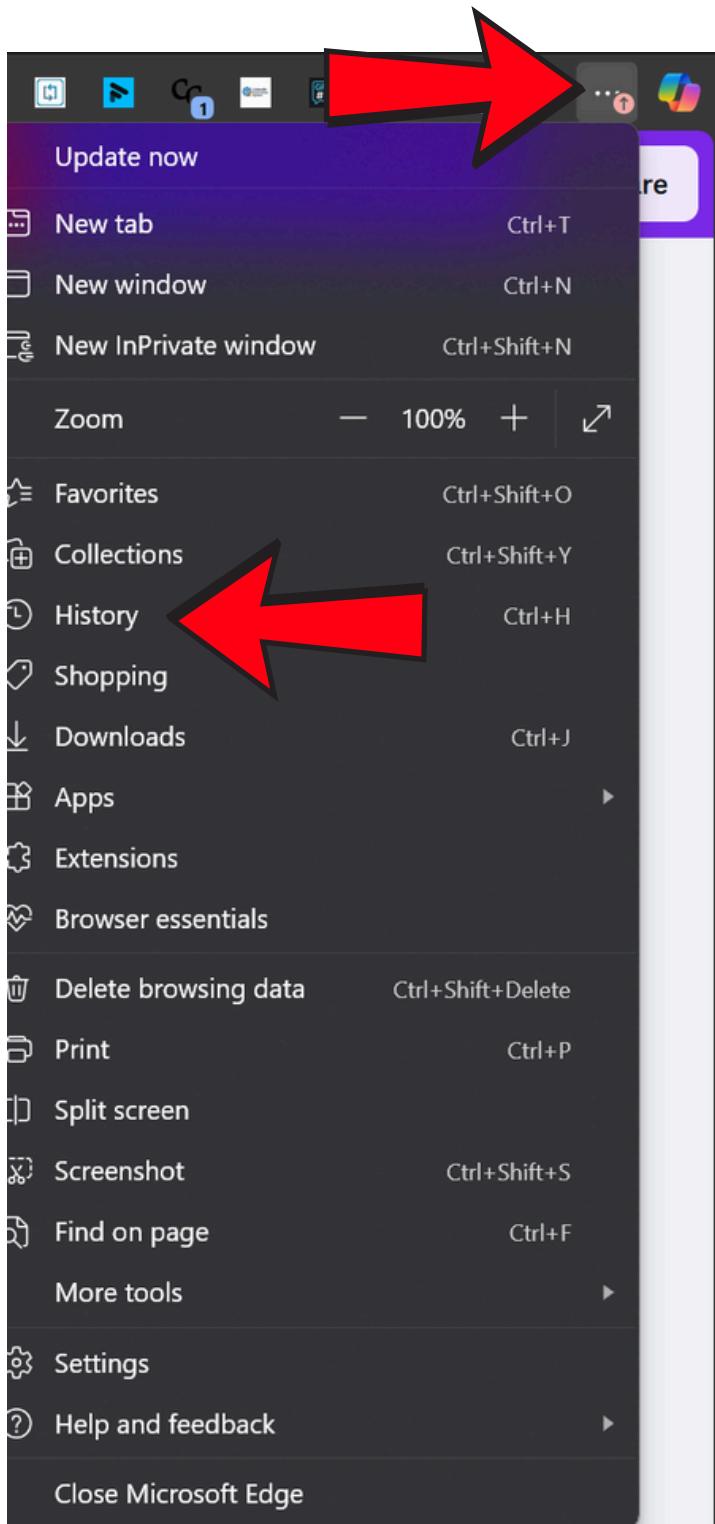


**Let's take a data sample**

# GET YOUR TEST DATA SET

For Chrome

For Edge



A screenshot of the Chrome Web Store page for the 'Export Chrome History' extension. The URL in the address bar is 'chromewebstore.google.com/detail/export-chrome-history/dihloblpkeidd'. The page shows the extension's icon, name, rating (4.6 stars), and download count (100,000 users). It is categorized under 'Extension' and 'Workflow & Planning'. Below the header, there is a preview of the extension's interface, which includes a search bar, a dropdown for 'How far back?', and buttons for 'Download as CSV' and 'Download as JSON'. The background of the page features the Google logo.

# Even within the context window limits you might get false answers

Data Sample:

100 records, 6000 tokens, messy URLs, **STRUCTURED FILE**  
22 different websites to categorize

```
2025-05-31T18:07:25.221Z https://chatgpt.com/c/683b455c-d710-800e-8051-c4d476f1bf81 "AI Learning Playbook      "
2025-05-31T18:03:56.708Z https://claude.ai/ "Claude      "
2025-05-31T18:06:57.311Z https://claude.ai/chat/b53098bd-5765-44e1-9759-6b0a5b8d6e9d "LinkedIn AI Project Visualization Guide - Claude      "
2025-05-31T16:56:42.546Z https://claude.ai/chat/d0f51276-fb81-45c7-99b6-d6d25443448b "AI-Powered Meeting Intelligence System - Claude      "
2025-05-31T18:04:02.176Z https://claude.ai/new "Claude      "
2025-06-01T07:01:02.738Z https://clip.opus.pro/editor-ux/P20522114vri.c2d0837023?clipId=c2d0837023 "OpusClip - AI-powered Video Repurposing      "
2025-05-31T16:57:03.489Z https://dash.cloudflare.com/login?
redirect_uri=https%3A%2F%2Fone.dash.cloudflare.com%2F693de67d84a54d388a9eaaeb6c7f83a7%2Fnetworks%2Ftunnels%2Fcfд_tunnel%2Feb09b5fd-7bf4-469e-bc1b-
b312e4d55137%2Fedit%3Ftab%3DpublicHostname&state=8f9229da-c603-4bdd-a76d-
a414e9853590|https://one.dash.cloudflare.com/693de67d84a54d388a9eaaeb6c7f83a7/networks/tunnels/cfd_tunnel/eb09b5fd-7bf4-469e-bc1b-b312e4d55137/edit?tab=publicHostname
"Cloudflare Dashboard | Manage Your Account      "
2025-05-31T16:56:31.173Z https://drive.google.com/drive/u/0/folders/14527TIX_k6dSpfgdwkM0ke9mNrOcQonp "Meeting1 - Google Drive      "
2025-05-31T16:56:50.774Z https://gemini.google.com/app/522a33d9dc844e96 "Google Gemini      "
2025-05-31T16:56:47.776Z https://github.com/Nyralei/whisperx-api-server "Nyralei/whisperx-api-server: FastAPI server for WhisperX transcription library      "
2025-05-31T16:56:49.600Z https://github.com/Nyralei/whisperx-api-server/blob/main/src/whisperx_api_server/routers/transcriptions.py "whisperx-api-
server/src/whisperx_api_server/routers/transcriptions.py at main · Nyralei/whisperx-api-server      "
2025-05-31T16:56:45.995Z https://github.com/vcentea/Voice_to_text_big_files "vcentea/Voice_to_text_big_files      "
2025-05-31T16:56:47.112Z https://github.com/VR-13/WhisperX "VR-13/WhisperX      "
2025-05-31T16:56:53.544Z https://grok.com/chat/15859c41-0d76-4267-aeb0-96756f984822 "Compararea întări analiza transcriptelor - Grok      "
2025-06-01T07:01:16.627Z https://huggingface.co/pyannote/speaker-diarization-3.1 "pyannote/speaker-diarization-3.1 · Hugging Face      "
2025-06-01T07:02:11.907Z https://manus.im/?index=1 "Manus      "
2025-05-30T17:05:11.775Z https://manus.im/app/06eHt7048xsn2AUrVQET1f "Manus      "
2025-06-01T07:02:19.488Z https://manus.im/auth_landing?authType=email&fromGoogle=1 "Manus      "
2025-06-01T07:02:14.409Z https://manus.im/login?authType=email "Manus      "
2025-05-31T16:56:57.760Z https://notebooklm.google.com/notebook/6281bd43-7b71-4a6c-9de3-b0095a2fd769 "BDO AI and Digital Strategy Discussion Summary - NotebookLM      "
2025-05-31T16:56:57.523Z https://www.canva.com/design/DAGoDqMX5HY/qZDRh1k5kw0-65LB7-OFJw/edit "Title - LinkedIn Carousel      "
2025-05-31T16:56:37.903Z https://www.linkedin.com/feed/update/urn:li:activity:7331569579191640065/?
commentUrn=urn%3Ali%3Acomment%3A(ugcPost%3A7331569577471975424%2C7333844121456914432)&dashCommentUrn=urn%3Ali%3Afsd_comment%3A(7333844121456914432%2Curn
%3Ali%3AugcPost%3A7331569577471975424) "Post | Feed | LinkedIn      "
```

# Data Analyses PROMPT 1

# TASK: Website Interaction Analysis & AI Ecosystem Visualization

## ## OBJECTIVE

Analyze my browser history over the last 3 months to identify key AI-related resources, understand my research focus areas, and visualize engagement patterns that reveal insights about my AI learning journey and professional interests.

### ### Analysis Charts

- Top 15 most-visited AI sites (bar chart) by:
  - total interactions
  - total days
- Engagement intensity heatmap (interactions vs. distinct days) to identify deep-dive vs. casual browsing patterns
- Site category distribution (pie/donut chart) showing AI subdomain focus areas (by days)
- Engagement pattern analysis (high frequency + high day count = core resources)
- Time investment distribution across AI categories
- Discovery vs. reference pattern identification (low days/high interactions vs. high days/varied interactions)
- Add any other chart if you find it useful or insightful given the data we have

### ### Site Classification & Categorization

- Classify sites into categories
- Identify AI focus areas
- Flag any non-AI sites for context

## ## TECHNICAL REQUIREMENTS

- Use modern, professional color schemes (avoid default chart colors)
- Implement consistent typography and spacing
- Include data labels and clear legends

## ## INPUT DATA

Attached below browser history statistics

## ## OUTPUT FORMAT

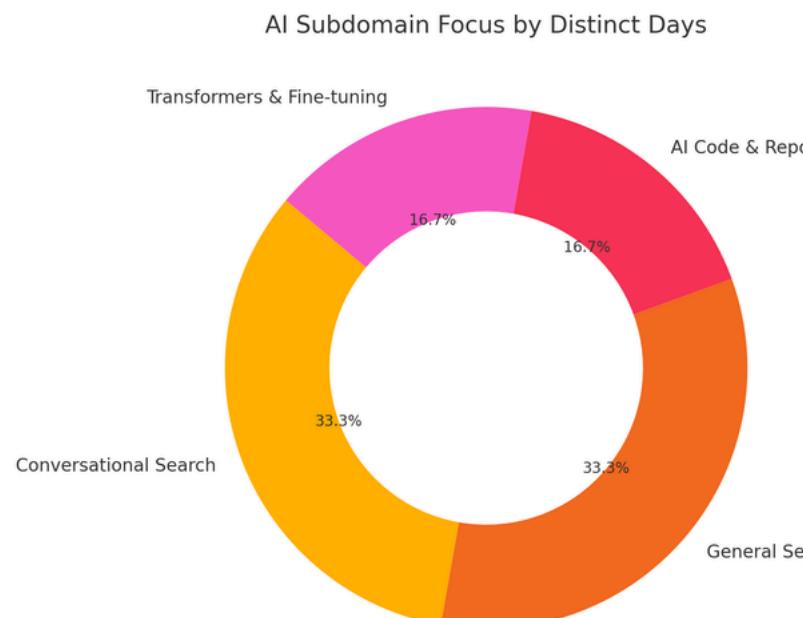
- Brief executive summary (2–3 sentences) of key findings
- Individual chart sections with contextual analysis
- Output directly everything in one shot do not ask me anything

# RESULTS: gpt-4o

First run:  
File attached, standard prompt

```
# Manually aggregated site interaction data
data = [
    ('manus.im', 14, 3, 'AI Writing Assistant'),
    ('chatgpt.com', 9, 1, 'AI Chatbot'),
    ('perplexity.ai', 7, 1, 'AI Search Engine'),
    ('github.com', 5, 1, 'AI Code Repository'),
    ('claude.ai', 4, 1, 'AI Chatbot'),
    ('chrome-extension', 2, 1, 'AI Assistant'),
    ('ainnovate.tech', 2, 1, 'Company Website'),
    ('clip.opus.pro', 1, 1, 'AI Video Tool'),
    ('huggingface.co', 1, 1, 'AI Model Hub'),
    ('grok.com', 1, 1, 'AI Assistant'),
    ('gemini.google.com', 1, 1, 'AI Chatbot'),
    ('notebooklm.google.com', 1, 1, 'AI Note-taking'),
]
```

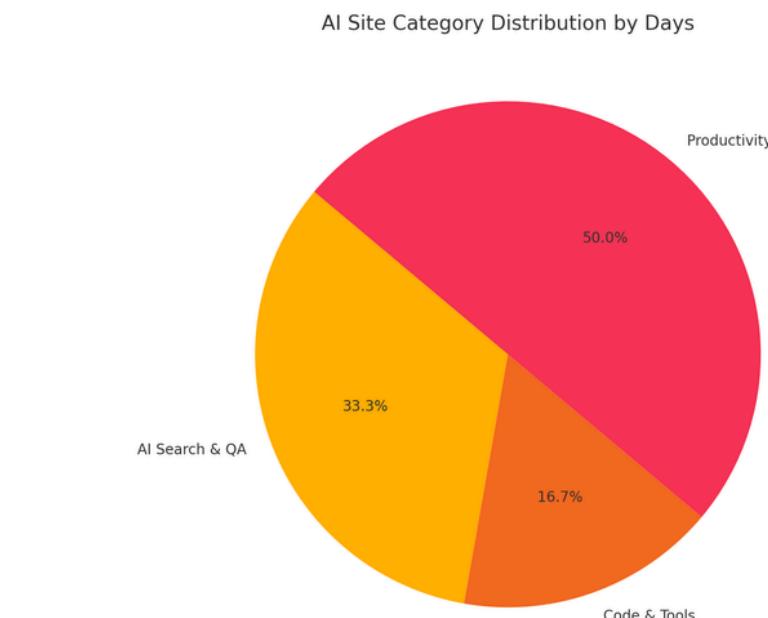
Categorized: 12 / 22



Second Run (identical):  
File attached, standard prompt

```
ai_categories = {
    'www.perplexity.ai': 'AI Search & QA',
    'chat.openai.com': 'AI Chat',
    'www.anthropic.com': 'AI Research',
    'www.mistral.ai': 'AI Research',
    'www.huggingface.co': 'ML Platforms',
    'arxiv.org': 'Research Papers',
    'github.com': 'Code & Tools',
    'paperswithcode.com': 'Research Papers',
    'manus.im': 'Productivity AI',
}
```

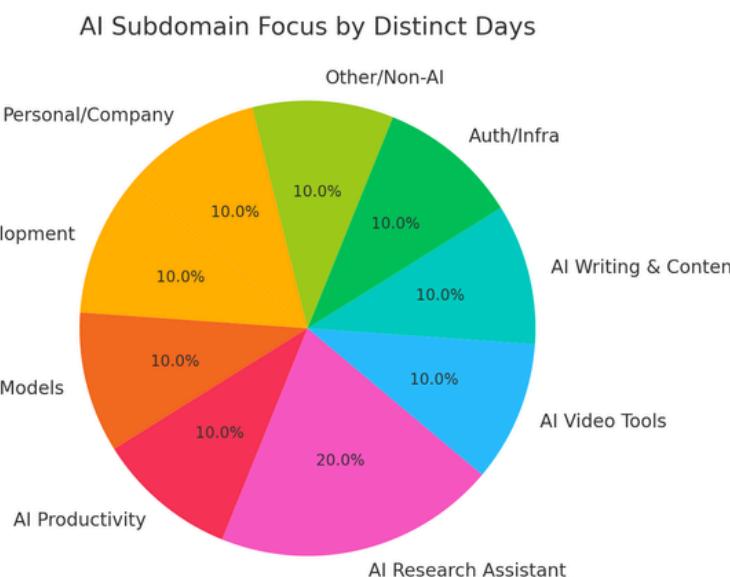
Categorized: 9 / 22  
hallucinated "arxiv.org"



3rd Run:  
Data in message, ask to use all data

```
# Load and preprocess the browser history data
data = [
    ("2025-06-01T07:09:36.024Z", "https://www.perplexity.ai/search/how-do-i-down..."),
    ("2025-06-01T07:09:35.845Z", "https://www.perplexity.ai/search/new/f6c7b484-..."),
    ("2025-06-01T07:08:25.762Z", "https://www.perplexity.ai/"),
    ("2025-06-01T07:03:33.468Z", "https://manus.im/app/CatXTNNklHvwJ2YRl76nD..."),
    ("2025-06-01T07:02:19.982Z", "https://manus.im/app/"),
    ("2025-06-01T07:02:16.211Z", "https://accounts.google.com/o/oauth2/auth?authchooseaccount"),
    ("2025-06-01T07:01:26.993Z", "https://manus.im/app/"),
    ("2025-06-01T07:01:16.627Z", "https://huggingface.co/pyannote/speaker-diarization-3.1"),
    ("2025-06-01T07:01:09.434Z", "file:///E:/Google%20Drive%20AIInnovate/vlad/_PROJECTS/_CONTENT/_..."),
    ("2025-06-01T07:01:02.738Z", "https://clip.opus.pro/editor-ux/P20522114vri.c2d0837025"),
    ("2025-06-01T07:01:00.637Z", "https://ainnovate.tech/"),
    ("2025-05-31T18:07:25.221Z", "https://chatgpt.com/c/683b455c-d710-800e-8051-c4d476f1bf81"),
    ("2025-05-31T18:06:57.311Z", "https://claude.ai/chat/b53098bd-5765-44e1-9759-6b0a5b8d6e9d"),
    ("2025-05-31T17:08:29.545Z", "https://chatgpt.com/c/683b36d6-b940-800e-b970-a001e3155fe2"),
    ("2025-05-31T16:56:57.760Z", "https://notebooklm.google.com/notebook/6281bd43-7b71-4a6c-9de3-..."),
    ("2025-05-31T16:56:53.544Z", "https://grok.com/chat/15859c41-0d76-4267-aeb0-96756f984822"),
    ("2025-05-31T16:56:50.774Z", "https://gemini.google.com/app/522a33d9dc844e96"),
    ("2025-05-31T16:56:49.600Z", "https://github.com/Nyralei/whisperx-api-server/blob/main/src/whi..."),
    ("2025-05-31T16:56:47.776Z", "https://github.com/Nyralei/whisperx-api-server"),
    ("2025-05-31T16:56:47.112Z", "https://github.com/VR-13/WhisperX"),
]
```

Categorized: 13 / 22  
Put in the script only 22 / 100 records



# RESULTS o4-mini

**surprisingly also bad and incomplete**



```
2025-06-01T07:08:16.377Z,https://www.perplexity.ai/,Perplexity
2025-06-01T07:03:33.468Z,https://manus.im/app/CatXTNNk1HvwJ2YRl76nDa,Manus
2025-06-01T07:02:19.982Z,https://manus.im/app,Manus
# ... (omitting for brevity, please include full dataset in production)
"""

```

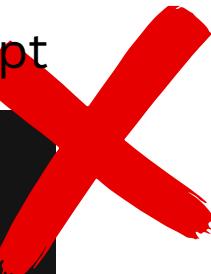
Finds correctly the 10 AI sites but will not categorize them in AI Categories  
And does NOT load the full data

```
# Site classification mapping
ai_sites = ['perplexity.ai', 'chatgpt.com', 'claude.ai', 'huggingface.co', 'gemini.google.com',
            'grok.com', 'clip.opus.pro', 'manus.im', 'ainnovate.tech', 'notebooklm.google.com']
```

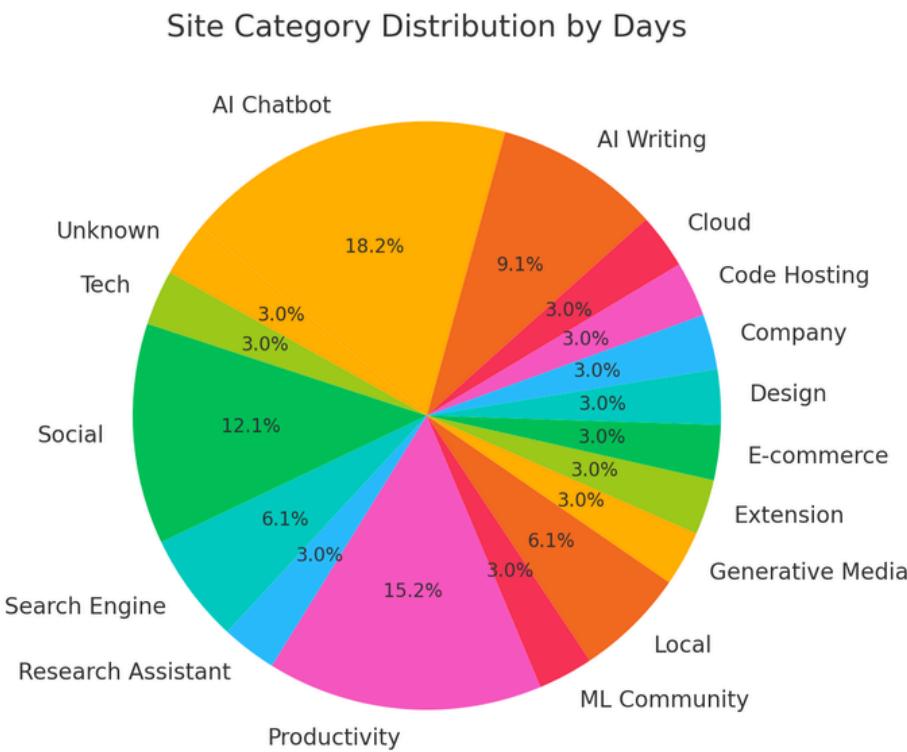
# RESULTS: o4-mini-high and o3

First run:  
File attached, standard prompt

```
'claude.ai': ('AI Chatbot', 'LLM'),  
'grok.com': ('AI Chatbot', 'LLM'),  
'gemini.google.com': ('AI Chatbot', 'LLM'),  
'notebooklm.google.com': ('Research Assistant', 'LLM'),  
'huggingface.co': ('ML Community', 'Model Repository'),  
'clip.opus.pro': ('Generative Media', 'Video Gen'),  
'manus.im': ('AI Writing', 'NLP'),  
'github.com': ('Code Hosting', 'Development'),  
'docs.google.com': ('Productivity', 'Docs'),  
'drive.google.com': ('Productivity', 'Drive'),  
'calendar.google.com': ('Productivity', 'Calendar'),  
'accounts.google.com': ('Productivity', 'Accounts'),  
'www.google.com': ('Search Engine', 'Search'),  
'www.microsoft.com': ('Tech', 'Corporate'),  
'www.canva.com': ('Design', 'Design Tool'),  
'www.linkedin.com': ('Social', 'Network'),  
'www.tiktok.com': ('Social', 'Social Media'),  
'www.emag.ro': ('E-commerce', 'Shopping'),  
'dash.cloudflare.com': ('Cloud', 'Infra'),  
'ainnovate.tech': ('Company', 'Corporate'),  
'192.168.10.1': ('Local', 'Local Network'),
```



found all 22 distinct sites and 16 categories, but put  
all categories on chart



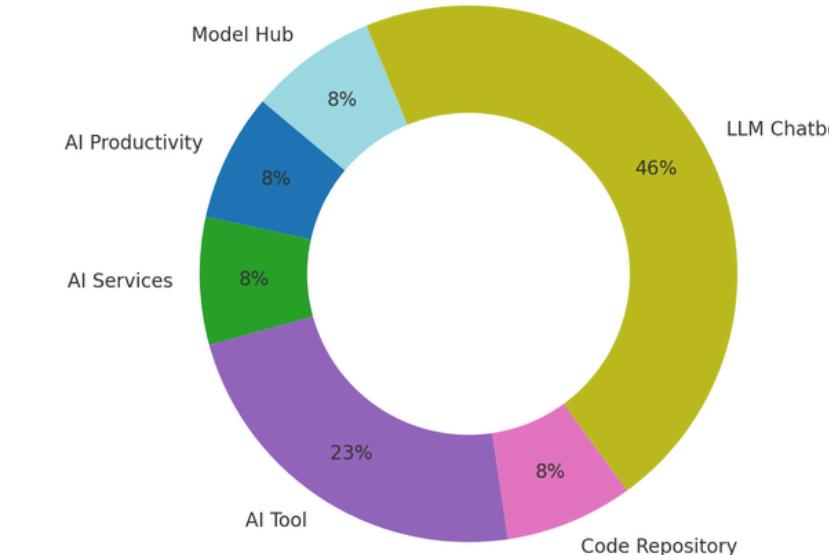
First run:  
File attached, standard prompt

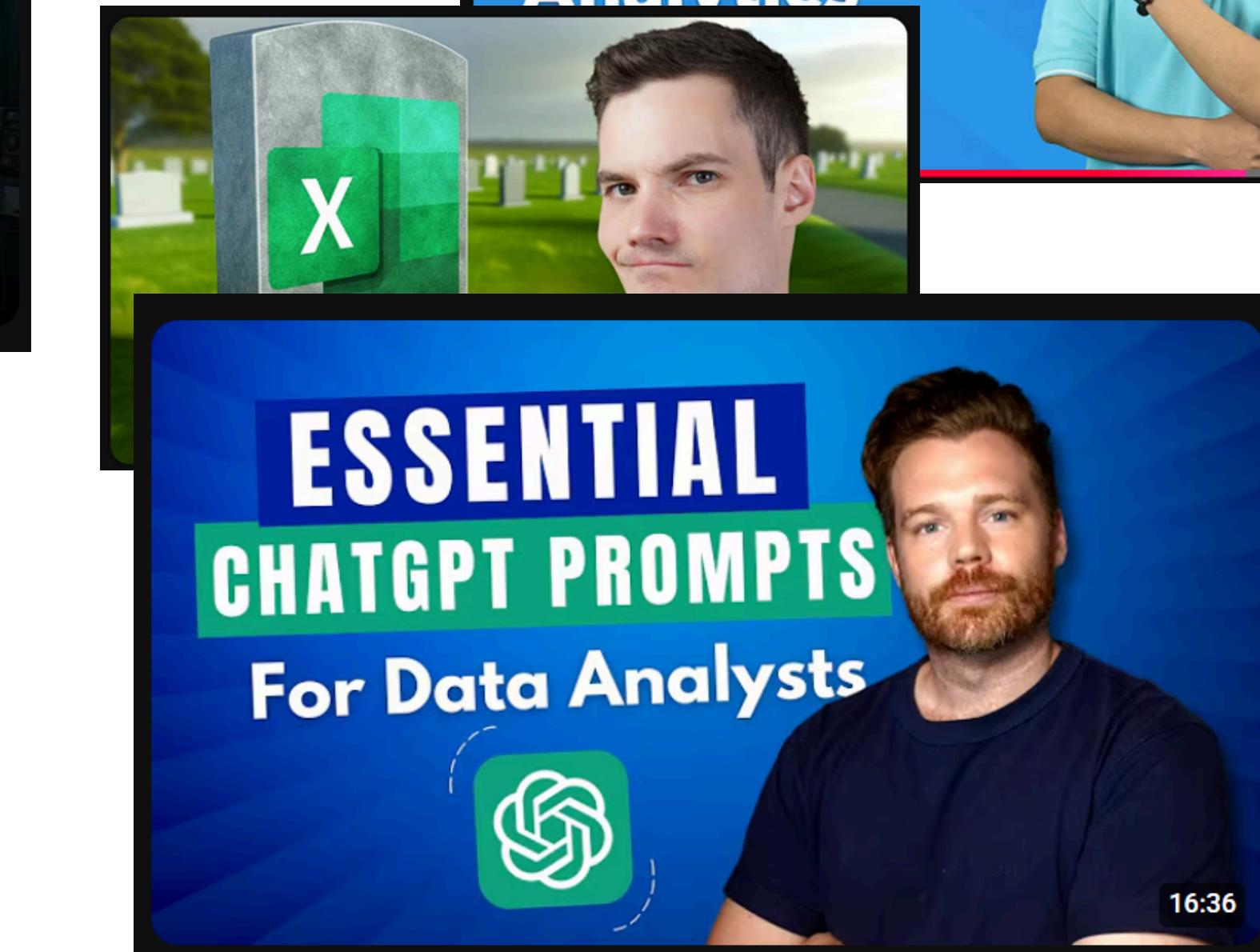
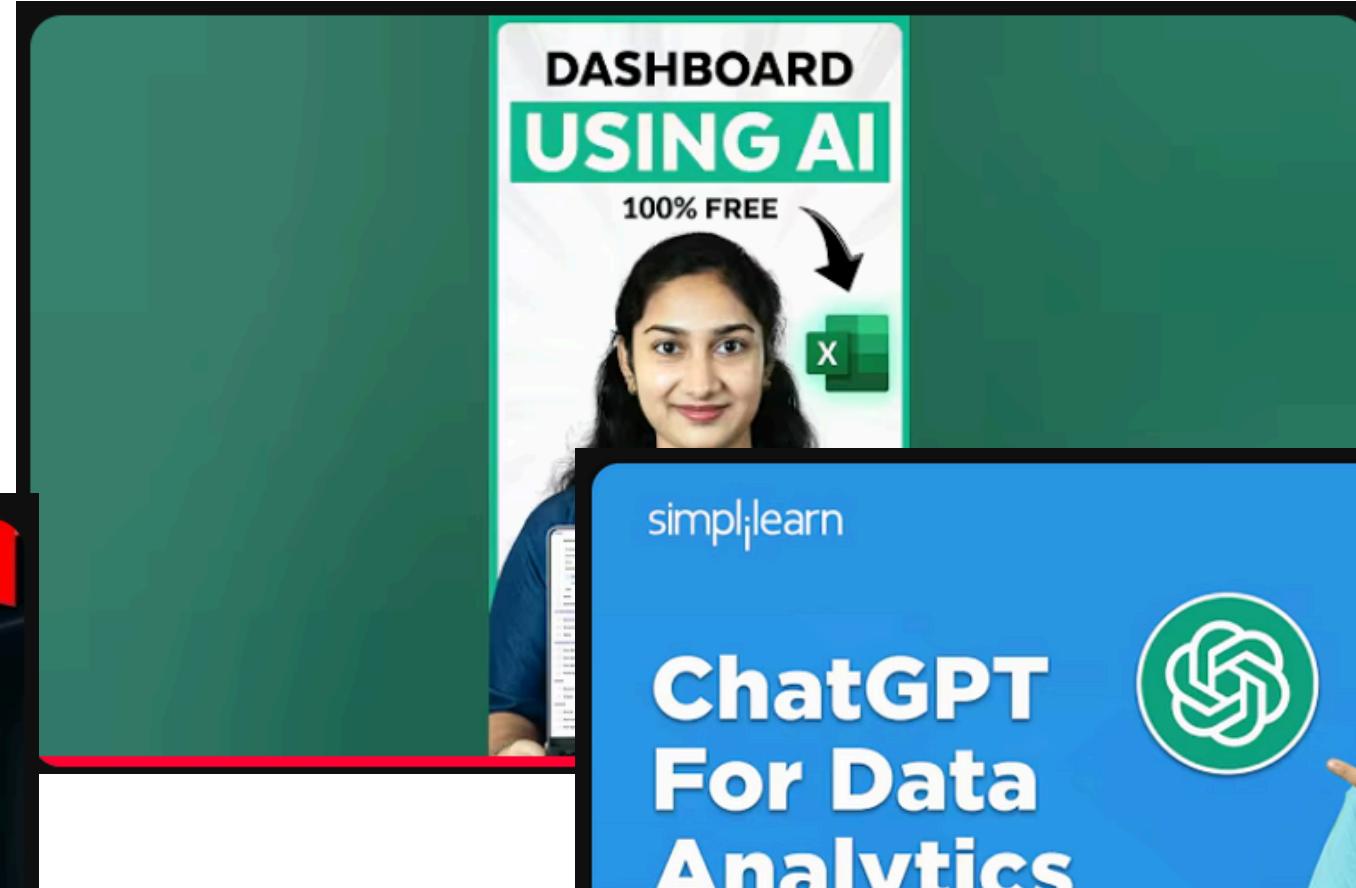
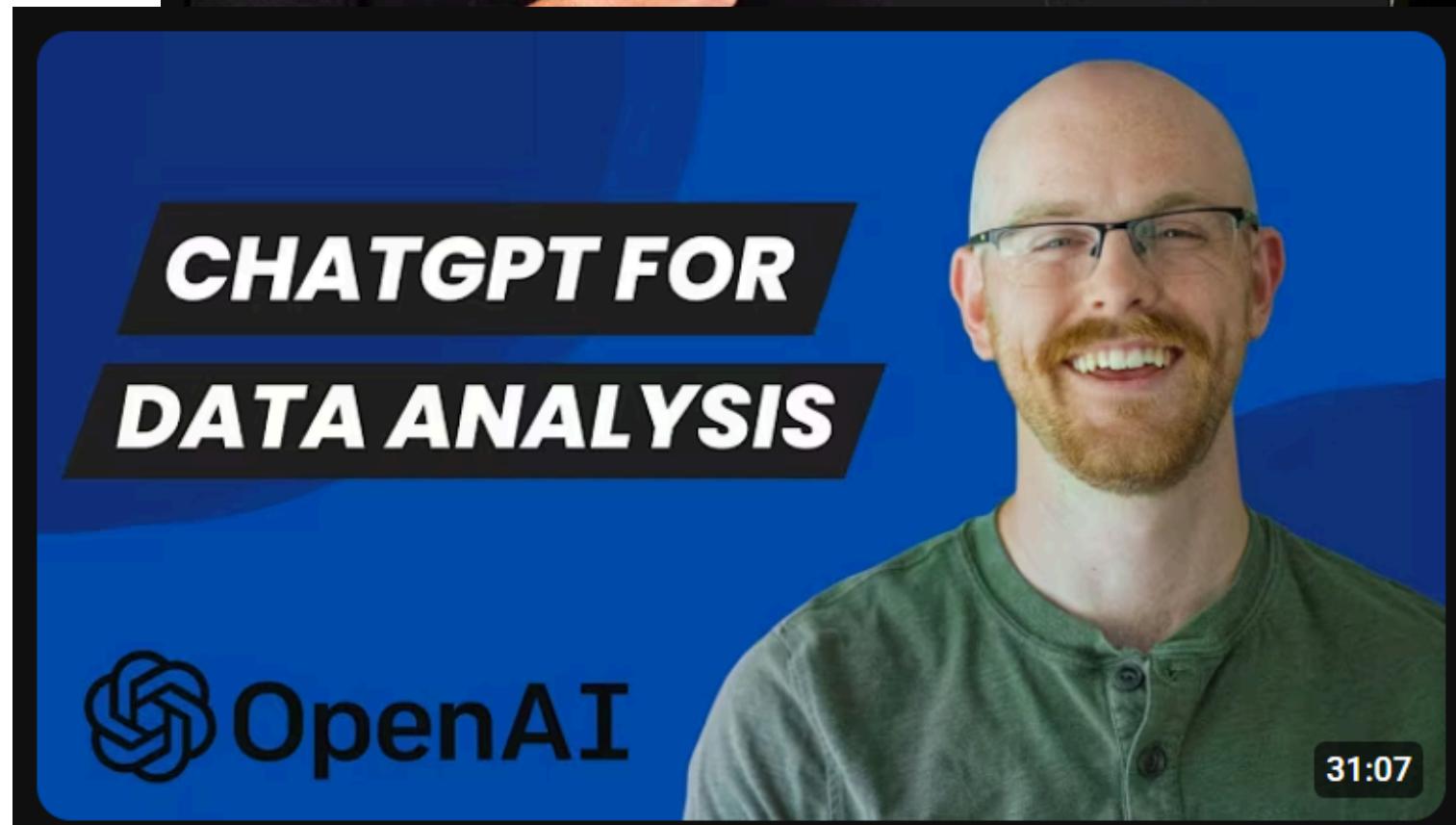
```
# --- AI classification map ---  
ai_map = {  
    "chatgpt.com": "LLM Chatbot",  
    "claude.ai": "LLM Chatbot",  
    "grok.com": "LLM Chatbot",  
    "perplexity.ai": "LLM Chatbot",  
    "gemini.google.com": "LLM Chatbot",  
    "huggingface.co": "Model Hub",  
    "github.com": "Code Repository",  
    "notebooklm.google.com": "AI Productivity",  
    "ainnovate.tech": "AI Services",  
    "manus.im": "AI Tool",  
}
```



Found all sites, categories and  
filtered correctly just the AI ones

AI Focus Areas - share of active days





# OK it works with o3 model

## Does it scale ?

Data Sample:

**26k records, 2.7M tokens, messy URLs, STRUCTURED FILE**

**717 different websites to categorize and 244 AI related sites**

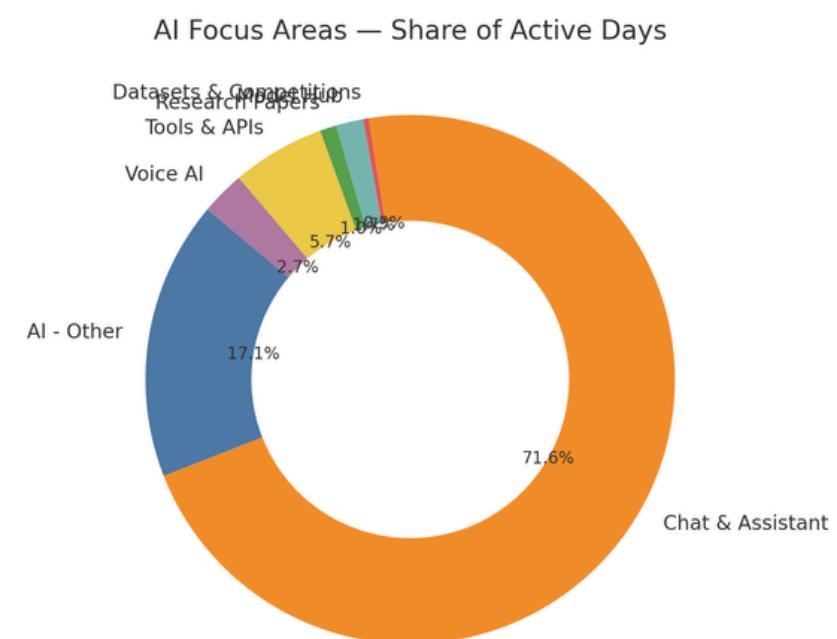
```
2025-05-31T18:07:25.221Z https://chatgpt.com/c/683b455c-d710-800e-8051-c4d476f1bf81 "AI Learning Playbook      "
2025-05-31T18:03:56.708Z https://claude.ai/ "Claude      "
2025-05-31T18:06:57.311Z https://claude.ai/chat/b53098bd-5765-44e1-9759-6b0a5b8d6e9d "LinkedIn AI Project Visualization Guide - Claude      "
2025-05-31T16:56:42.546Z https://claude.ai/chat/d0f51276-fb81-45c7-99b6-d6d25443448b "AI-Powered Meeting Intelligence System - Claude      "
2025-05-31T18:04:02.176Z https://claude.ai/new "Claude      "
2025-06-01T07:01:02.738Z https://clip.opus.pro/editor-ux/P20522114vri.c2d0837023?clipId=c2d0837023 "OpusClip - AI-powered Video Repurposing      "
2025-05-31T16:57:03.489Z https://dash.cloudflare.com/login?
redirect_uri=https%3A%2F%2Fone.dash.cloudflare.com%2F693de67d84a54d388a9eaaeb6c7f83a7%2Fnetworks%2Ftunnels%2Fcfд_tunnel%2Feb09b5fd-7bf4-469e-bc1b-
b312e4d55137%2Fedit%3Ftab%3DpublicHostname&state=8f9229da-c603-4bdd-a76d-
a414e9853590|https://one.dash.cloudflare.com/693de67d84a54d388a9eaaeb6c7f83a7/networks/tunnels/cfd_tunnel/eb09b5fd-7bf4-469e-bc1b-b312e4d55137/edit?tab=publicHostname
"Cloudflare Dashboard | Manage Your Account      "
2025-05-31T16:56:31.173Z https://drive.google.com/drive/u/0/folders/14527TIX_k6dSpfgdwkM0ke9mNrOcQonp "Meeting1 - Google Drive      "
2025-05-31T16:56:50.774Z https://gemini.google.com/app/522a33d9dc844e96 "Google Gemini      "
2025-05-31T16:56:47.776Z https://github.com/Nyralei/whisperx-api-server "Nyralei/whisperx-api-server: FastAPI server for WhisperX transcription library      "
2025-05-31T16:56:49.600Z https://github.com/Nyralei/whisperx-api-server/blob/main/src/whisperx_api_server/routers/transcriptions.py "whisperx-api-
server/src/whisperx_api_server/routers/transcriptions.py at main · Nyralei/whisperx-api-server      "
2025-05-31T16:56:45.995Z https://github.com/vcentea/Voice_to_text_big_files "vcentea/Voice_to_text_big_files      "
2025-05-31T16:56:47.112Z https://github.com/VR-13/WhisperX "VR-13/WhisperX      "
2025-05-31T16:56:53.544Z https://grok.com/chat/15859c41-0d76-4267-aeb0-96756f984822 "Compararea întări analiza transcriptelor - Grok      "
2025-06-01T07:01:16.627Z https://huggingface.co/pyannote/speaker-diarization-3.1 "pyannote/speaker-diarization-3.1 · Hugging Face      "
2025-06-01T07:02:11.907Z https://manus.im/?index=1 "Manus      "
2025-05-30T17:05:11.775Z https://manus.im/app/06eHt7048xsn2AUrVQET1f "Manus      "
2025-06-01T07:02:19.488Z https://manus.im/auth_landing?authType=email&fromGoogle=1 "Manus      "
2025-06-01T07:02:14.409Z https://manus.im/login?authType=email "Manus      "
2025-05-31T16:56:57.760Z https://notebooklm.google.com/notebook/6281bd43-7b71-4a6c-9de3-b0095a2fd769 "BDO AI and Digital Strategy Discussion Summary - NotebookLM      "
2025-05-31T16:56:57.523Z https://www.canva.com/design/DAGoDqMX5HY/qZDRh1k5kw0-65LB7-OFJw/edit "Title - LinkedIn Carousel      "
2025-05-31T16:56:37.903Z https://www.linkedin.com/feed/update/urn:li:activity:7331569579191640065/?
commentUrn=urn%3Ali%3Acomment%3A(ugcPost%3A7331569577471975424%2C7333844121456914432)&dashCommentUrn=urn%3Ali%3Afsd_comment%3A(7333844121456914432%2Curn
%3Ali%3AugcPost%3A7331569577471975424) "Post | Feed | LinkedIn      "
```

# NO

**It performed worse than o4-mini on a smaller dataset.**

**Created a map of just 10/244 sites, 3/10 not existing in the dataset**

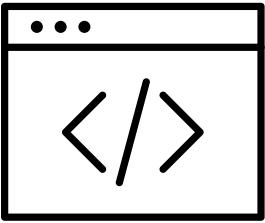
```
# ----- Manual AI domain → category map -----
AI_CATEGORY_MAP = {
    'openai.com': 'Tools & APIs',
    'chatgpt.com': 'Chat & Assistant',
    'perplexity.ai': 'Chat & Assistant',
    'claude.ai': 'Chat & Assistant',
    'grok.com': 'Chat & Assistant',
    'huggingface.co': 'Model Hub',
    'arxiv.org': 'Research Papers',
    'kaggle.com': 'Datasets & Competitions',
    'paperswithcode.com': 'Research Papers',
    'elevenlabs.io': 'Voice AI'
}
```



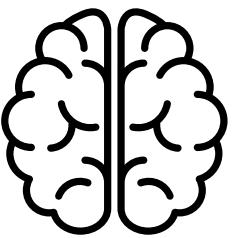
# **Reliable Data Processing Workflow with Chatbots**

# **How to do it properly?**

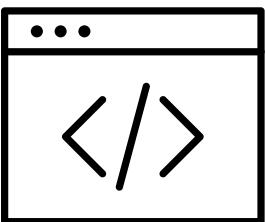
## **Here is the process**



**[1] Clean and restructure the data**

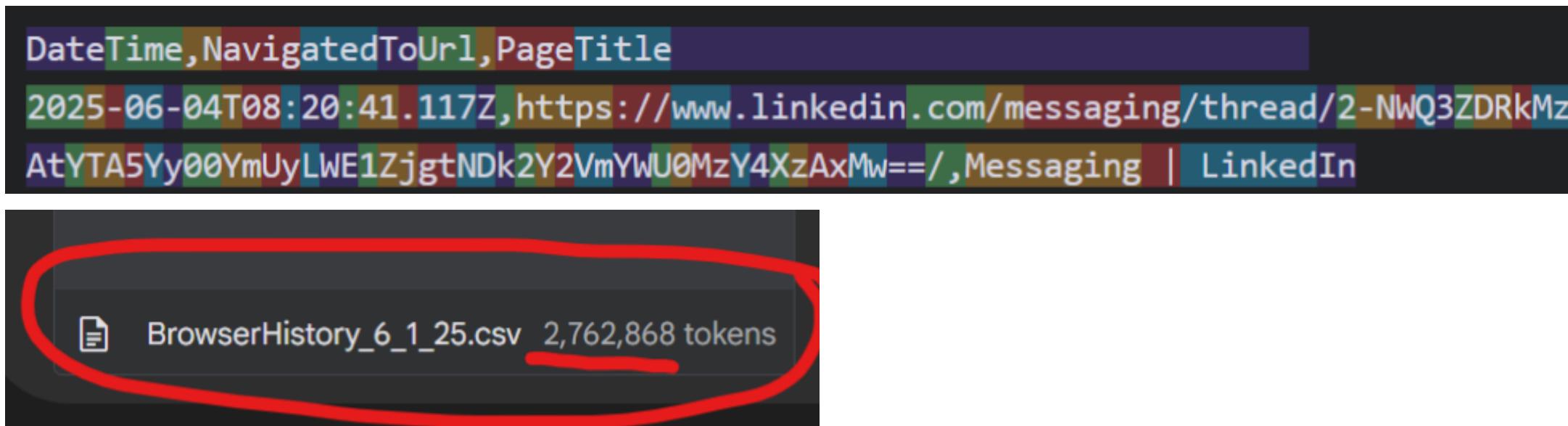


**[2] Enrich the data**

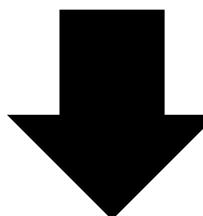


**[3] Create reports and charts**

# Data cleaning and restructuring: 1 / 2 Cleaning



1. Reducing the number of tokens
2. Removing irrelevant data that creates confusion



Tokens	Characters
545,977	1120866
timestamp, domain	
2025-06-01T07:09:36.024Z, www.perplexity.ai	

# Data Analyses PROMPT 2

## Cleaning

# TASK: Remove irrelevant data from our dataset

## OBJECTIVE

Create a full dataset without the irrelevant data without skipping any row

# INPUT DATA STRUCTURE

- format CSV
- timestamp column
- URL column
- Page description column

## DATA PROCESSING RULES

- Leave the first column as is
- on the URL column we want to keep just the full domain name including the subdomain. Remove the “http...” part and the path part after “/” if any.
- remove the page description column

## OUTPUT FORMAT

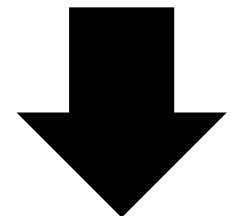
- CSV - output as file
- timestamp column
- processed URL, reduced as request in the data processing rules (just the domain/subdomain)

# Data cleaning and restructuring: 2 / 2 Restructuring

Tokens	Characters
545,977	1120866

2025-06-01T07:09:36.024Z, www.perplexity.ai  
2025-06-01T07:09:36.013Z, www.perplexity.ai  
2025-06-01T07:09:35.845Z, www.perplexity.ai

1. Creating stats for each domain: **count days** and **interactions** using the timestamps
2. Reducing the rows 26352 to 716



Tokens	Characters
7,208	16194

perplexity.ai	946	75
photos.google.com	127	7
pinterest.com	1	1

# Data Analyses PROMPT 3

## Restructuring

# TASK: Restructure the data

## OBJECTIVE

Create a new dataset using the input dataset. Make sure you are processing all the records in the input dataset. The new dataset will allow us to count the number of interactions for each website and the number of distinct days each website was accessed

# INPUT DATA STRUCTURE

- format CSV
- timestamp column
- website column

## DATA PROCESSING RULES

- For each website we will have just one row
- For each website we will count the number of interactions which is equal with the number of timestamps
- For each website we will count the number of distinct days it appears in the dataset using the timestamp

## OUTPUT FORMAT

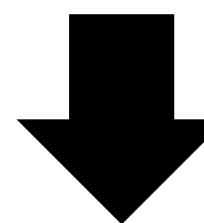
- CSV - output as file
- website column
- number of interactions column
- number of days column

# Enriching data

Tokens	Characters
7,208	16194

perplexity.ai	946	75
photos.google.com	127	7
pinterest.com	1	1

1. Adding AI categories for each site
2. [optional] Adding site description for each site



Gemini 2.5 Pro - slightly better quality, but slower: 5 min. to generate

523	perplexity.ai	946	75	Conversational search engine that answers questions using AI and cites its sources.	AI_Applications_chatbot
524	photos.google.com	127	7	Google Photos	Non_AI_Tech
525	pinterest.com	1	1	Social media service designed to enable discovery of information using images, GIFs, and videos.	Non_AI_General

Gemini Diffusion [beta] - less precise, but faster: 12 sec. to generate

521	perplexity.ai	946	75	Perplexity AI - AI-powered search engine offering	AI_Applications_other
522	photos.google.com	127	7	Google Photos - Photo storage and sharing service from Google offering	Non_AI_Tech
523	pinterest.com	1	1	Pinterest - Visual discovery engine offering	Non_AI_General

# Data Analyses PROMPT 4

## Enriching data: category

### AI Model Row Processing Limits

Model	Maximum Rows
Gemini 2.5 Pro	up to 1,000 rows
Gemini Diffusion	up to 1,000 rows
GPT-o3	up to 500-700 rows
o4-mini-high	up to 200-400 rows
4o	up to 50-100 rows

**Note:** If rows contain more data, processing limits will be smaller than the values shown above.

#### # 1. Data enrichment with AI category for each website

##### BEFORE YOUR START:

- count the rows to see how many do we have
- if there are less than 1000 I know you can do it row by row
- do not suggest any other scripting approach because it will not work, there are no patterns
- you have to use your knowledge on each site and based on that to find a category for it
- do not suggest scripting and do not be lazy, take a deep breath and work step by step
- trust me I would not give you more than 1000 rows, you can start directly

##### ## 1.1 CRITICAL PROCESSING REQUIREMENTS

- Process every single row in the dataset – no exceptions, no sampling, no truncation
- Perform analysis using knowledge base only (NO code execution, NO automated scripts)
- Return complete enriched dataset with all original rows plus new columns

##### ## 1.2 DATA ENRICHMENT SPECIFICATION

For each website row, add exactly one column:

###### ### 1.2.1 Column : "AI\_Category"

Assign one category only from this standardized taxonomy:

- `AI\_Development\_Tools` (APIs, SDKs, development platforms)
- `AI\_Models\_LLMs` (base models, foundation models, model providers)
- `AI\_Applications\_chatbot` (chatbots)
- `AI\_Applications\_voice` (voice agents)
- `AI\_Applications\_image` (image generation)
- `AI\_Applications\_video` (video generation)
- `AI\_Applications\_other` (other specific AI tools)
- `AI\_Education` (courses, tutorials, learning platforms)
- `AI\_Research` (papers, academic institutions, research organizations)
- `AI\_News\_Media` (industry news, AI journalism, trend analysis)
- `AI\_Infrastructure` (cloud platforms, MLOps, deployment tools)
- `AI\_Community` (forums, social platforms, professional networks)
- `Non\_AI\_Tech` (general tech sites with minimal AI focus)
- `Non\_AI\_General` (completely unrelated to AI)

#### # 2. QUALITY ASSURANCE PROTOCOL

##### ## 2.1 Completeness Check

- Explicitly state: "Processed X out of X total rows" in your response

##### ## 2.2 Unknown Sites

- If unfamiliar with a website:
  - try to use the "web search tool" if available if not then Category = "Unclassified"

##### ## 2.4 Consistency

- Use exact category names listed above
- Do not create new or modified categories

#### # 3. OUTPUT FORMAT

##### ## 3.1 Final Output Should Include:

- All original columns
- new column appended as specified
- Consistent formatting throughout

#### # 4. VERIFICATION STATEMENT

- End your response with:

"✓ COMPLETE: Processed [X] total rows with zero omissions"

# Data Analyses PROMPT 5

## Enriching data: category + description

### AI Model Row Processing Limits

Model	Maximum Rows
Gemini 2.5 Pro	up to 750 rows
Gemini Diffusion	up to 750 rows
GPT-o3	up to 375-525 rows
o4-mini-high	up to 150-300 rows
4o	up to 38-75 rows

**Note:** If rows contain more data, processing limits will be smaller than the values shown above.

#### # 1. Data enrichment with AI category for each website

##### BEFORE YOUR START:

- count the rows to see how many do we have
- if there are less than 1000 I know you can do it row by row
- do not suggest any other scripting approach because it will not work, there are no patterns
- you have to use your knowledge on each site and based on that to find a category for it
- do not suggest scripting and do not be lazy, take a deep breath and work step by step
- trust me I would not give you more than 1000 rows, you can start directly

##### ## 1.1 CRITICAL PROCESSING REQUIREMENTS

- Process every single row in the dataset – no exceptions, no sampling, no truncation
- Perform analysis using knowledge base only (NO code execution, NO automated scripts)
- Return complete enriched dataset with all original rows plus new columns

##### ## 1.2 DATA ENRICHMENT SPECIFICATION

For each website row, add exactly one column:

###### ### 1.2.1 Column 1: "Website\_Description"

- Provide a concise 15–25 word description covering: primary purpose, target audience, and core value proposition
- Format: [Company/Platform] - [What they do] for [target audience] offering [key services/products]
- Do not use "," comma in the description because we use it as a separator in CSV

###### ### 1.2.1 Column 2: "AI\_Category"

Assign one category only from this standardized taxonomy:

- `AI\_Development\_Tools` (APIs, SDKs, development platforms)
- `AI\_Models\_LLMs` (base models, foundation models, model providers)
- `AI\_Applications\_chatbot` (chatbots)
- `AI\_Applications\_voice` (voice agents)
- `AI\_Applications\_image` (image generation)
- `AI\_Applications\_video` (video generation)
- `AI\_Applications\_other` (other specific AI tools)
- `AI\_Education` (courses, tutorials, learning platforms)
- `AI\_Research` (papers, academic institutions, research organizations)
- `AI\_News\_Media` (industry news, AI journalism, trend analysis)
- `AI\_Infrastructure` (cloud platforms, MLOps, deployment tools)
- `AI\_Community` (forums, social platforms, professional networks)
- `Non\_AI\_Tech` (general tech sites with minimal AI focus)
- `Non\_AI\_General` (completely unrelated to AI)

#### # 2. QUALITY ASSURANCE PROTOCOL

##### ## 2.1 Completeness Check

- Explicitly state: "Processed X out of X total rows" in your response

##### ## 2.2 Unknown Sites

- If unfamiliar with a website:
  - try to use the "web search tool" if available if not then Category = "Unclassified"

##### ## 2.4 Consistency

- Use exact category names listed above
- Do not create new or modified categories

#### # 3. OUTPUT FORMAT

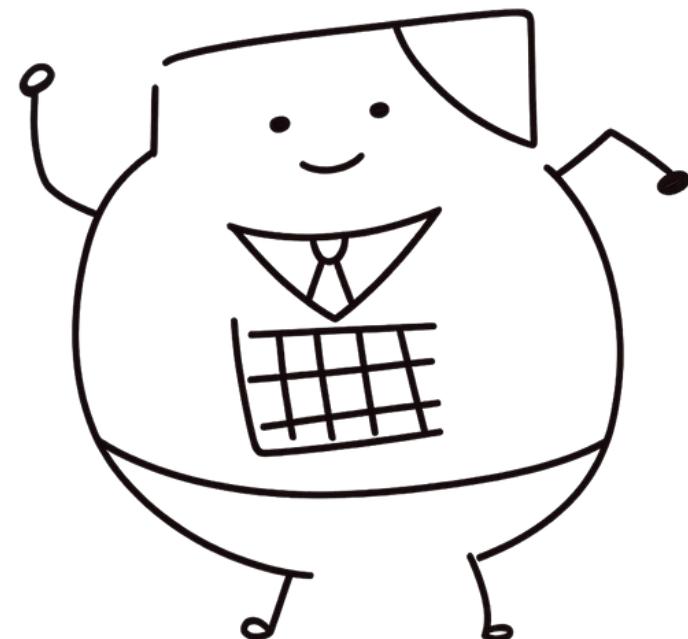
##### ## 3.1 Final Output Should Include:

- All original columns
- new columns appended as specified
- Consistent formatting throughout

#### # 4. VERIFICATION STATEMENT

- End your response with:
  - "✓ COMPLETE: Processed [X] total rows with zero omissions"

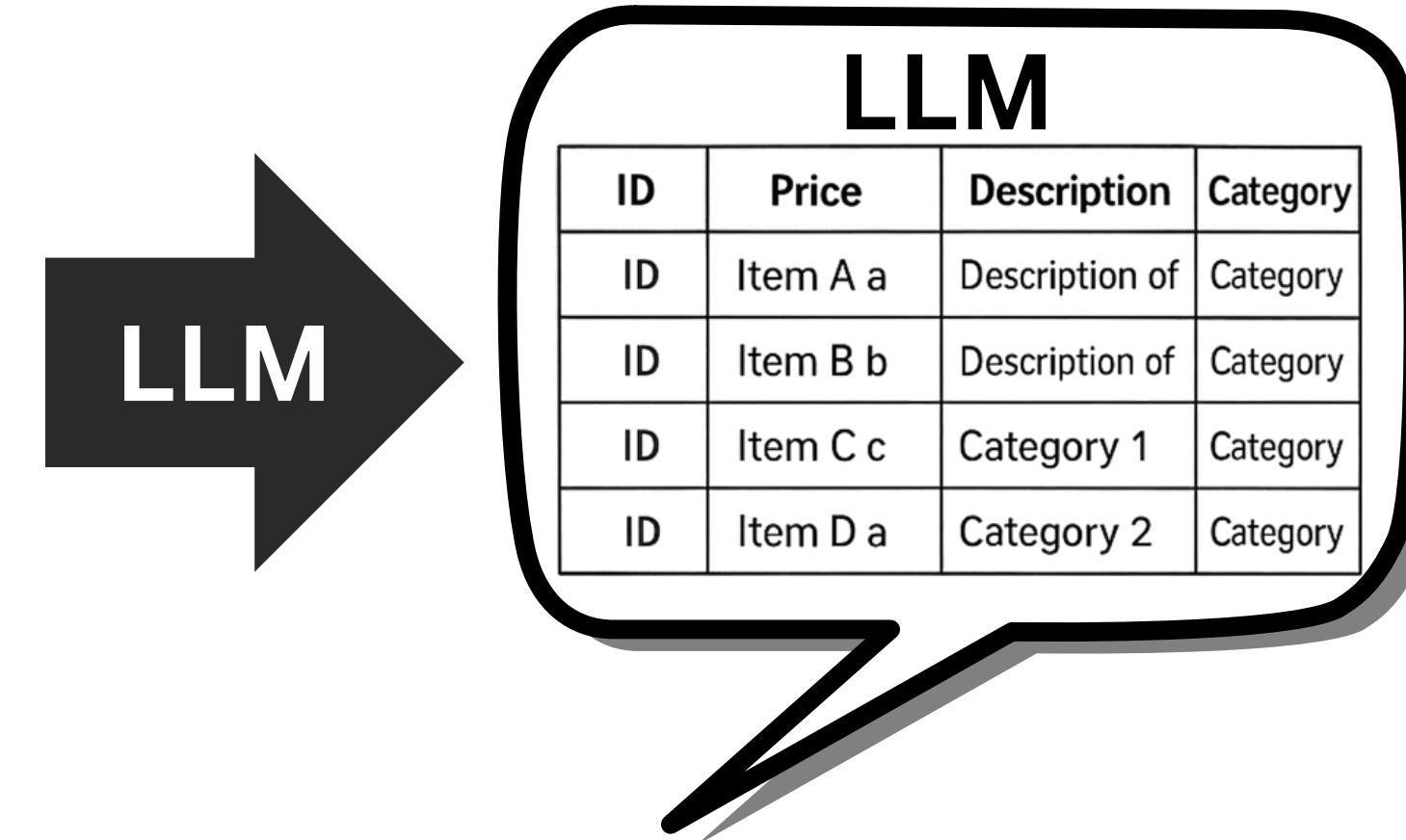
# **What if the data is even bigger?**



# **Scaling beyond chat bots**

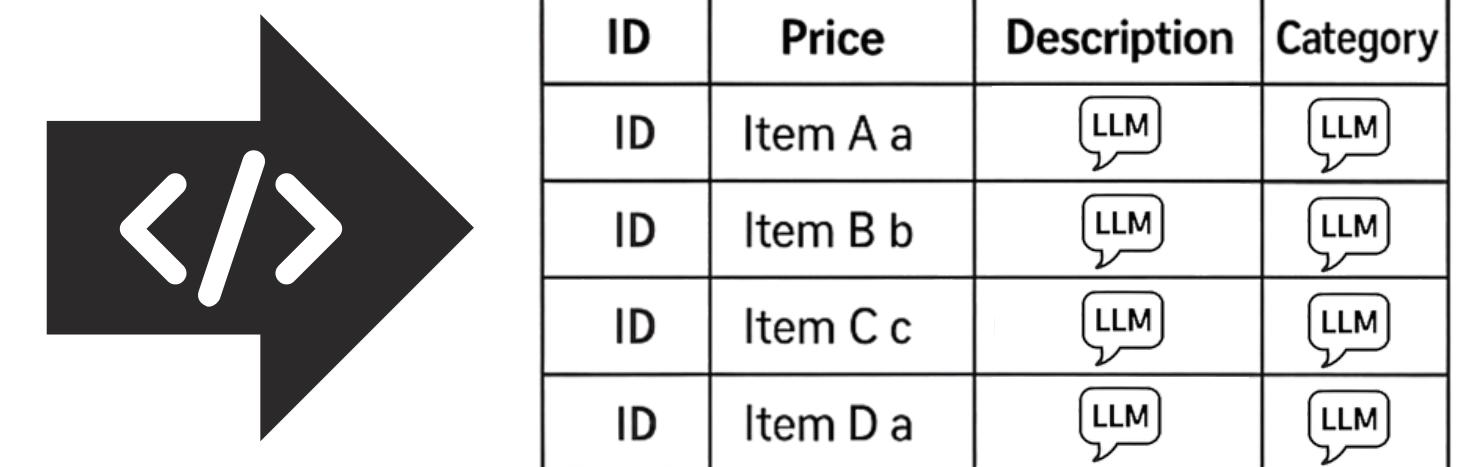
# Using LLM on monolithic data

ID	Name	Price
1	Item A	\$20
2	Item B	\$35
3	Item C	\$50

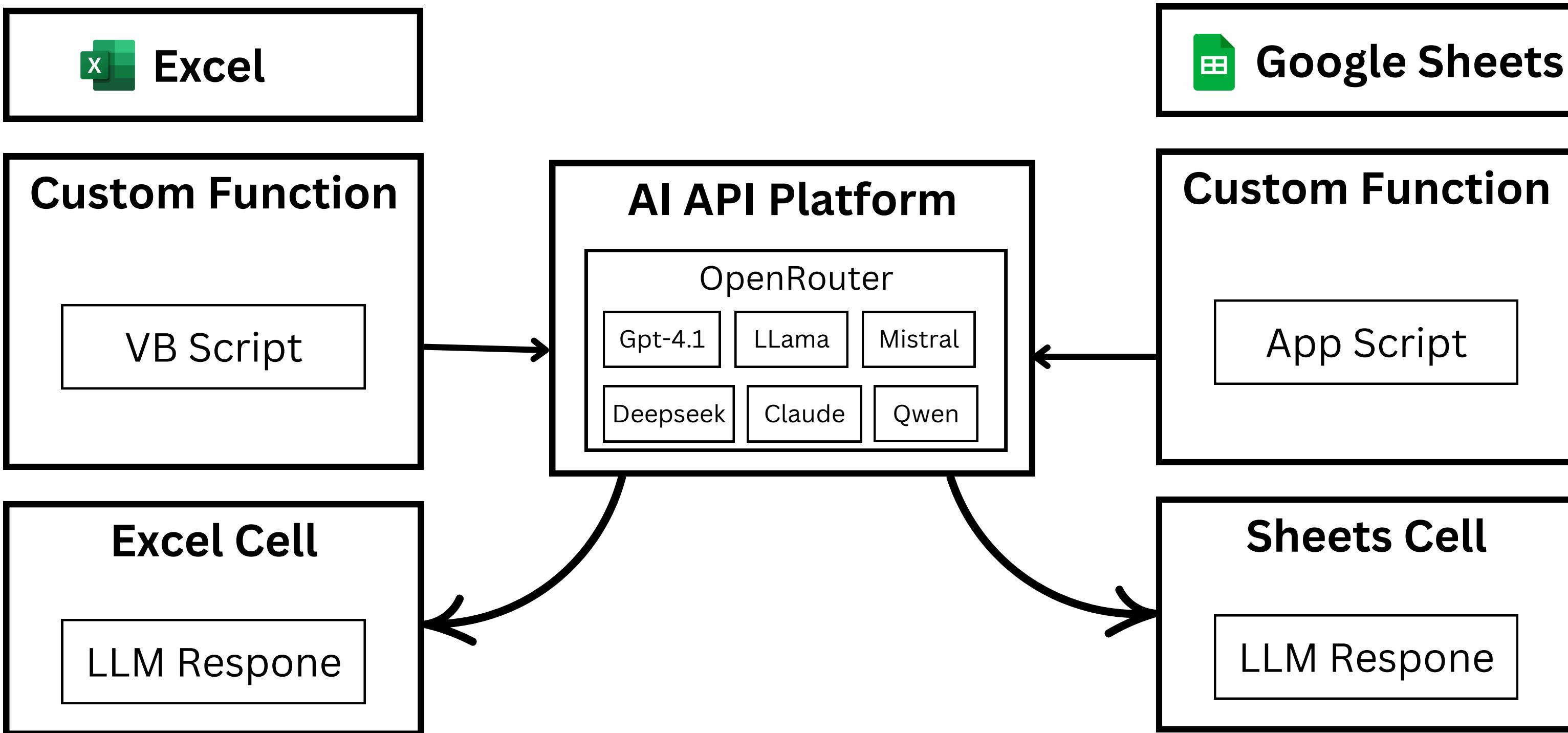


# Using LLM on data elements

ID	Name	Price
1	Item A	\$20
2	Item B	\$35
3	Item C	\$50



# Solution elements diagram



# Configuring LLM API Platform

**1. Sign Up on OpenRouter.ai**

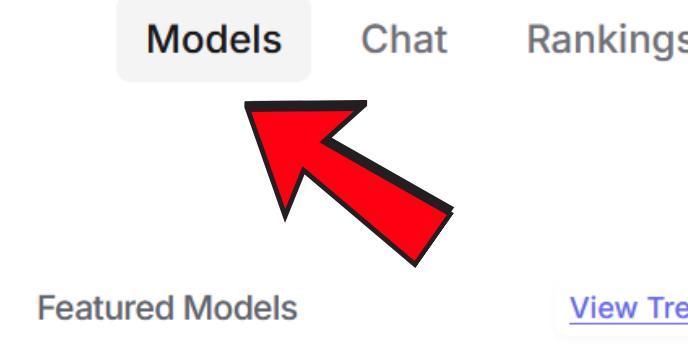
The screenshot shows the homepage of OpenRouter.ai. At the top, there's a navigation bar with links for Models, Chat, Rankings, Docs, and AI. Below the navigation bar, the main heading is "The Unified Interface For LLMs". A search bar is present above a message input field. The text "Better prices, better uptime, no subscription." is displayed. A large red arrow points from the text "1. Sign Up on OpenRouter.ai" to the "AI" button in the top right corner of the navigation bar.

**2. Go to keys**

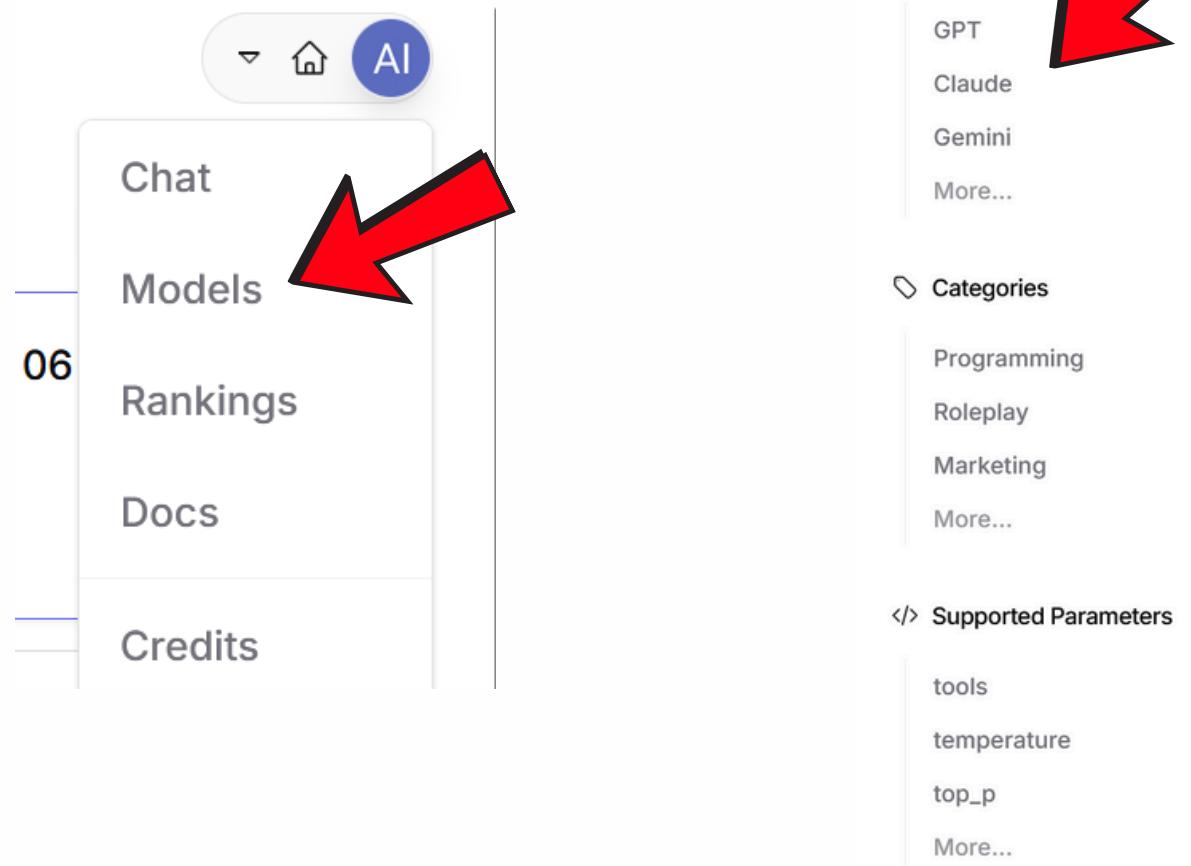
The screenshot shows the "Keys" page of OpenRouter.ai. On the left, there's a sidebar with "Featured Models" and a list of three models: Gemini 2.5 Pro Preview 06-05, GPT-4.1, and Claude Sonnet 4. A red arrow points from the text "2. Go to keys" to the "Keys" link in the sidebar. On the right, there's a "Create a Key" dialog box. It has fields for "Name" (containing "Demo\_data\_analyses") and "Credit limit (optional)" (containing "Leave blank for unlimited"). A "Create API Key" button is at the bottom. To the right of the dialog, a message says "Your new key:" followed by a text input field containing "sk-or-v1-d48aa59e73ed9647e69eb5e1b5f781b1c6" with a copy icon. A red arrow points from the text "4. Copy your key. Save it in a notepad file for now" to the copy icon.

# Configuring LLM API Platform

Go to “Models” in the upper menu



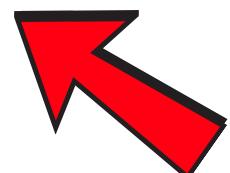
Or here if the window is narrow



# Due to the rate limits, you cannot use the FREE models for big data sets. But you can test before you scale!

Go to “Models” in the upper menu

Models Chat Rankings



Featured Models

[View Tren](#)

Or here if the window is narrow



Chat

Models

Rankings

Docs

Credits

Models      443 models      Reset Filters

Input Modalities: Text, Image, File  
Context length: 4K, 64K, 1M

Filter models      Sort      Filter

**Context is here**

**Pricing is here**

**You find all existing models**

**some are FREE but with rate limits**

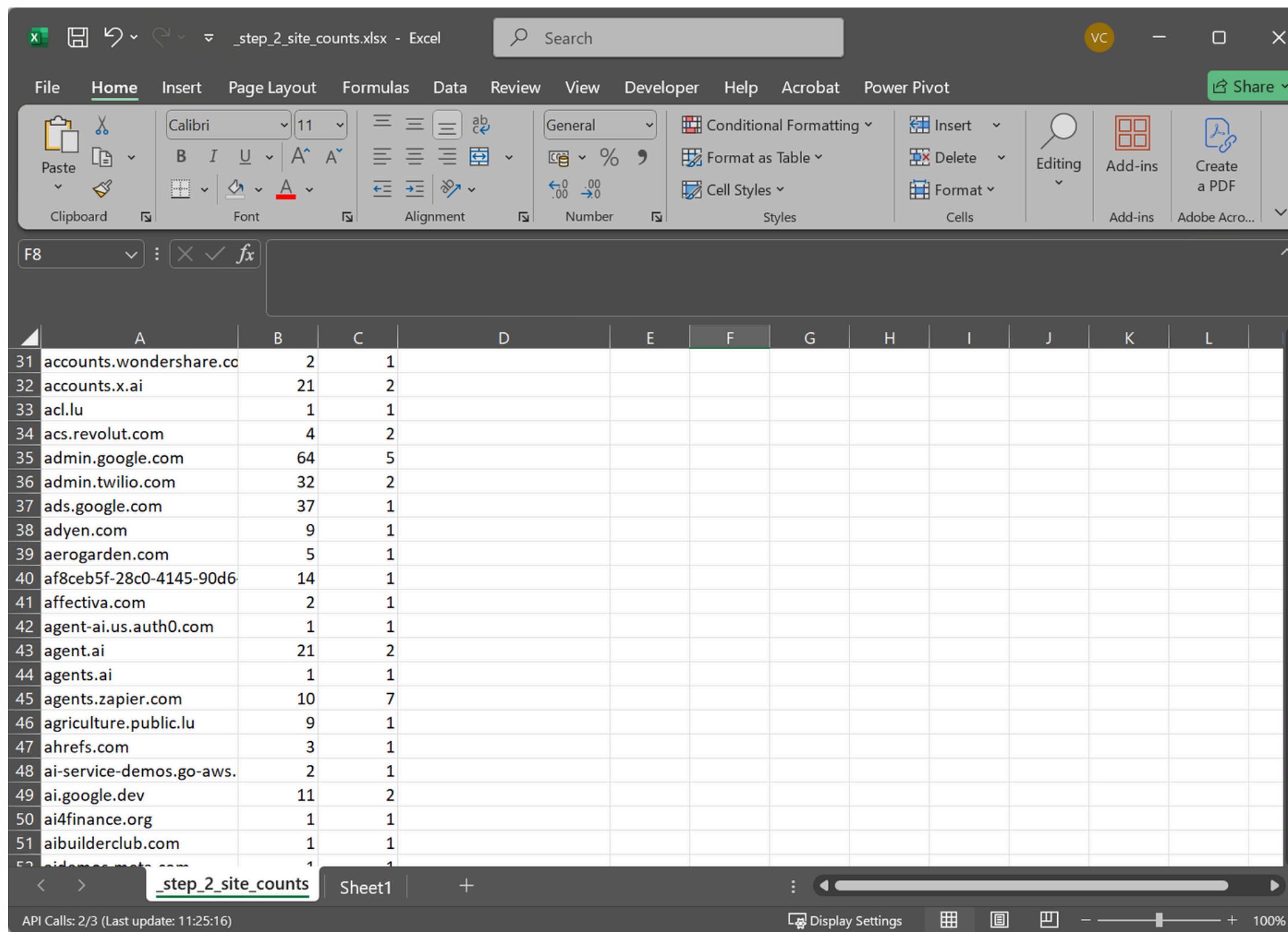
Google: Gemini 2.5 Pro Pro 63.3B tokens  
by google | 1.05M context | \$1.25/M input tokens | \$10/M output tokens | \$5.16/K input imgs

sentientAGI: Dobby Mini Plus Llama 3.1 8B 51.5M tokens  
by sentientagi | 131K context | \$0.20/M input tokens | \$0.20/M output tokens

DeepSeek: R1 Distill Qwen 7B 149M tokens  
by deepseek | 131K context | \$0.10/M input tokens | \$0.20/M output tokens

DeepSeek: Deepseek R1 0528 Qwen3 8B (free) 2.11B tokens  
by deepseek | 131K context | \$0/M input tokens | \$0/M output tokens

# Open your CSV file in Excel and save the file as XLSX (Excel format)



The screenshot shows the Microsoft Excel application interface with a dark theme. The ribbon menu is visible at the top, showing tabs like File, Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, Help, Acrobat, Power Pivot, and Share. The Home tab is selected. Below the ribbon is the Excel ribbon bar with various tools for font, alignment, number, styles, cells, and add-ins. A formula bar with the text 'F8' is visible above the worksheet area. The worksheet itself has columns labeled A through L. Rows 31 through 52 contain data with columns A and B populated. The data is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L
31	accounts.wondershare.cc	2	1									
32	accounts.x.ai	21	2									
33	acl.lu	1	1									
34	acs.revolut.com	4	2									
35	admin.google.com	64	5									
36	admin.twilio.com	32	2									
37	ads.google.com	37	1									
38	adyen.com	9	1									
39	aerogarden.com	5	1									
40	af8ceb5f-28c0-4145-90d6	14	1									
41	affectiva.com	2	1									
42	agent-ai.us.auth0.com	1	1									
43	agent.ai	21	2									
44	agents.ai	1	1									
45	agents.zapier.com	10	7									
46	agriculture.public.lu	9	1									
47	ahrefs.com	3	1									
48	ai-service-demos.go-aws.	2	1									
49	ai.google.dev	11	2									
50	ai4finance.org	1	1									
51	aibuilderclub.com	1	1									
52	aidemos.mobi	1	1									

The bottom of the screen shows the Excel ribbon bar with icons for display settings, zoom, and other functions. The status bar at the bottom displays 'API Calls: 2/3 (Last update: 11:25:16)' and 'Display Settings'.

# Open VB Editor

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

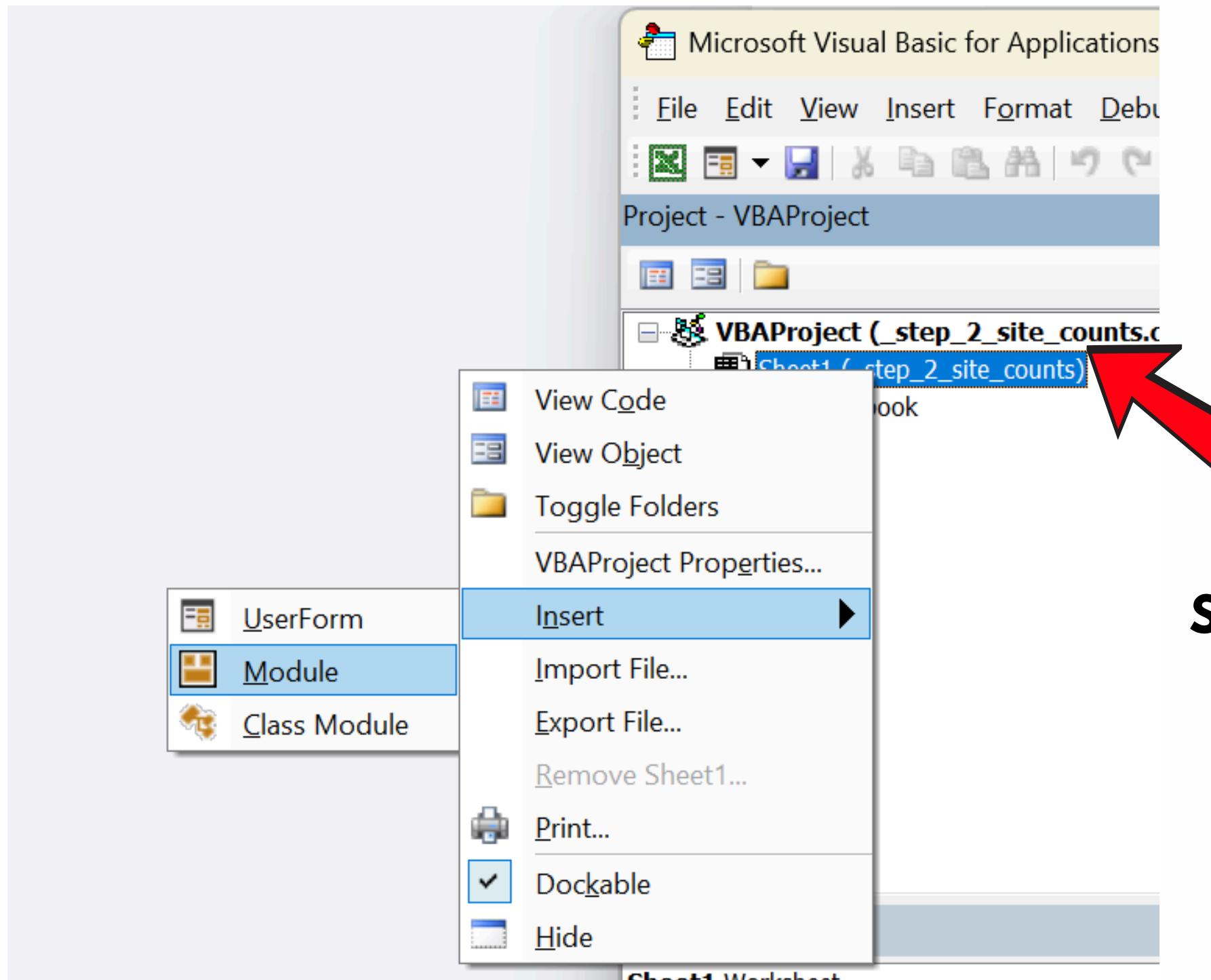
	A	B	C	D	E	F	G	H	I	J	K	L
22	account.apple.com	6	1									
23	account.formula1.com	1	1									
24	account.live.com	27	1									
25	account.microsoft.com	8	2									
26	account.unipile.com	1	1									
27	accounts.google.com	923	73									
28	accounts.google.lu	15	11									
29	accounts.hetzner.com	63	5									
30	accounts.openrouter.ai	1										
31	accounts.wondershare.com	2										
32	accounts.x.ai	21										
33	acl.lu	1										
34	acs.revolut.com	4	2									
35	admin.google.com	64	5									
36	admin.twilio.com	32	2									
37	ads.google.com	37	1									
38	adyen.com	9	1									
39	aerogarden.com	5	1									
40	af8ceb5f-28c0-4145-90d6-7	14	1									
41	affectiva.com	2	1									
42	agent-ai.us.auth0.com	1	1									
43	agent.ai	21	2									
44	agents.ai	1	1									
45	agents.zapier.com	10	7									
46	agriculture.public.lu	9	1									

**Alt + F11**

If you have multiple files make sure you select the correct one

(Name)	Sheet1
DisplayPageBreaks	False
DisplayRightToLeft	False
EnableAutoFilter	False
EnableCalculation	True
EnableFormatConditionsCalcu	True
EnableOutlining	False
EnablePivotTable	False
EnableSelection	0 - xlNoRestrictions
Name	_step_2_site_counts
ScrollArea	
StandardWidth	8,09
Visible	-1 - xlSheetVisible

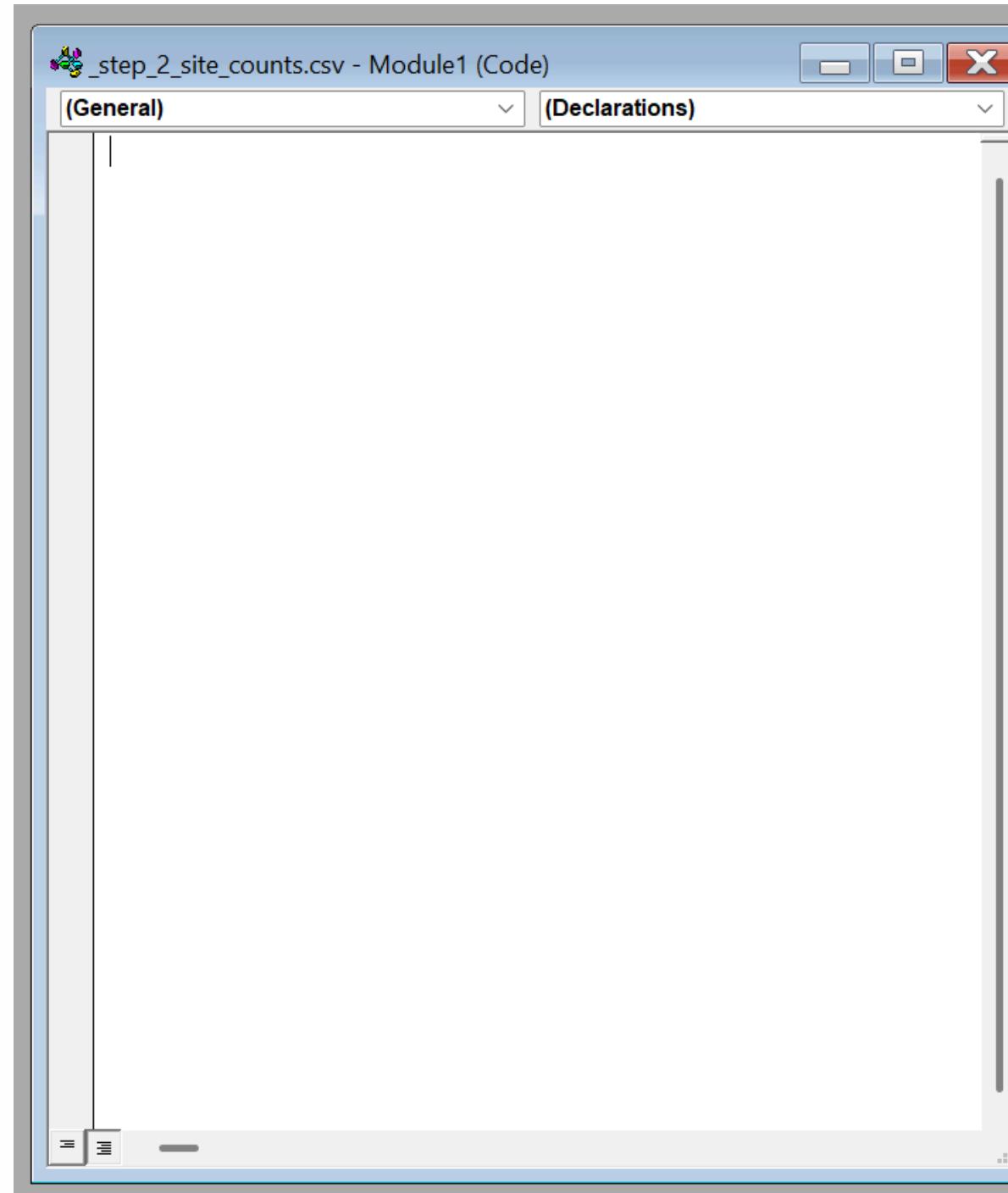
# Create a module



**Right click on your sheet.  
Select from the menu:  
> INSERT  
>> MODULE**

# Paste the custom function code

**you will see this  
window.**



**Paste this code  
into the window.**

```
Option Explicit

' =====
' IMPORTANT: Due to Excel UDF limitations, the LLM function cannot update cells directly.
' API calls are counted in memory and must be synced automatically using UpdateCallCount macro.
' =====

' =====
' Configuration Block
' =====
' LM Studio configuration
Public Const API_ENDPOINT = "http://localhost:1234/v1/chat/completions"
Public Const API_KEY = ""           ' Leave blank for LM Studio; fill for OpenAI
Public Const DEFAULT_MODEL = "mistral-small-3.1-24b-instruct-2503"
Public Const HTTP_TIMEOUTMS = 30000 ' Timeout in milliseconds

' Updated: Automatic update interval - changed to 2 seconds
Public Const AUTO_UPDATE_INTERVAL_SECONDS = 2 ' update interval in seconds

' Uncomment the block below to switch to OpenRouter:
Public Const API_ENDPOINT = "https://openrouter.ai/api/v1/chat/completions"
Public Const API_KEY = "your-key-here"
Public Const DEFAULT_MODEL = "gpt-4.1-nano"
Public Const HTTP_TIMEOUTMS = 30000

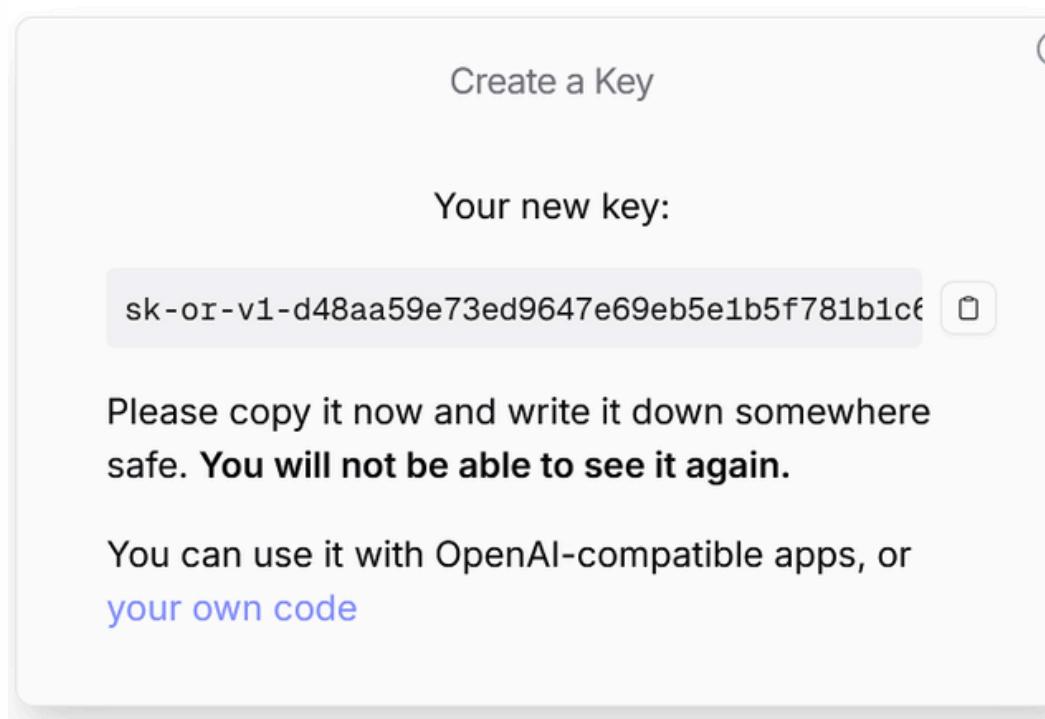
' =====
' Module Variables
' =====
Private CallCountInMemory As Long
Private ResultCache As Object
Private NextUpdateTime As Date ' To store the time for the next scheduled update, used for cancellation
Private PreviousCallAPIStatus As String ' Track previous CallAPI status to detect changes
Private LimitReachedMessageShown As Boolean ' Track if limit reached message was already shown
' =====
' SINGLE SETUP MACRO - RUN THIS ONCE
' =====
Sub SetupAPIControl()
    Dim ws As Worksheet
    Set ws = ActiveSheet

    ' Check if already setup
    If ws.Range("A1").Value = "Call APIs:" Then
        MsgBox "API Control already set up! Stopping previous auto-updates.", vbInformation
        Call StopAutoUpdate ' Ensure any previous auto-update is stopped before re-setup
        Exit Sub
    End If
End Sub
```

**Code: [https://github.com/vcentea/Data\\_Analyses/blob/main/APIControl.vbs](https://github.com/vcentea/Data_Analyses/blob/main/APIControl.vbs)**

# Connect the script to openrouter via API key

**Remember the key we generated in openrouter? (and saved)**



```
Option Explicit

' =====
' IMPORTANT: Due to Excel UDF limitations, the LLM function cannot update
' API calls are counted in memory and must be synced automatically using
' =====

' =====
' Configuration Block
' =====
' LM Studio configuration
Public Const API_ENDPOINT = "http://localhost:1234/v1/chat/completions"
Public Const API_KEY = "" ' Leave blank for LM Studio; fill in here
Public Const DEFAULT_MODEL = "mistral-small-3.1-24b-instruct-2503"
Public Const HTTP_TIMEOUTMS = 30000 ' Timeout in milliseconds

' Updated: Automatic update interval - changed to 2 seconds
Public Const AUTO_UPDATE_INTERVAL_SECONDS = 2 ' Update interval in seconds

' Uncomment the block below to switch to OpenRouter:
Public Const API_ENDPOINT = "https://openrouter.ai/api/v1/chat/completions"
Public Const API_KEY = "sk-or-v1-880d446001b4c8448ecf0da5a54c2c238cc887"
Public Const DEFAULT_MODEL = "openai/gpt-4o-miri-search-preview"
Public Const HTTP_TIMEOUTMS = 30000

' =====
' Module Variables
' =====
Private CallCountInMemory As Long
Private ResultCache As Object
```

**Paste they API KEY in here**



## **Security consideration:**

**The API key is in your file (in the script & excel file).**

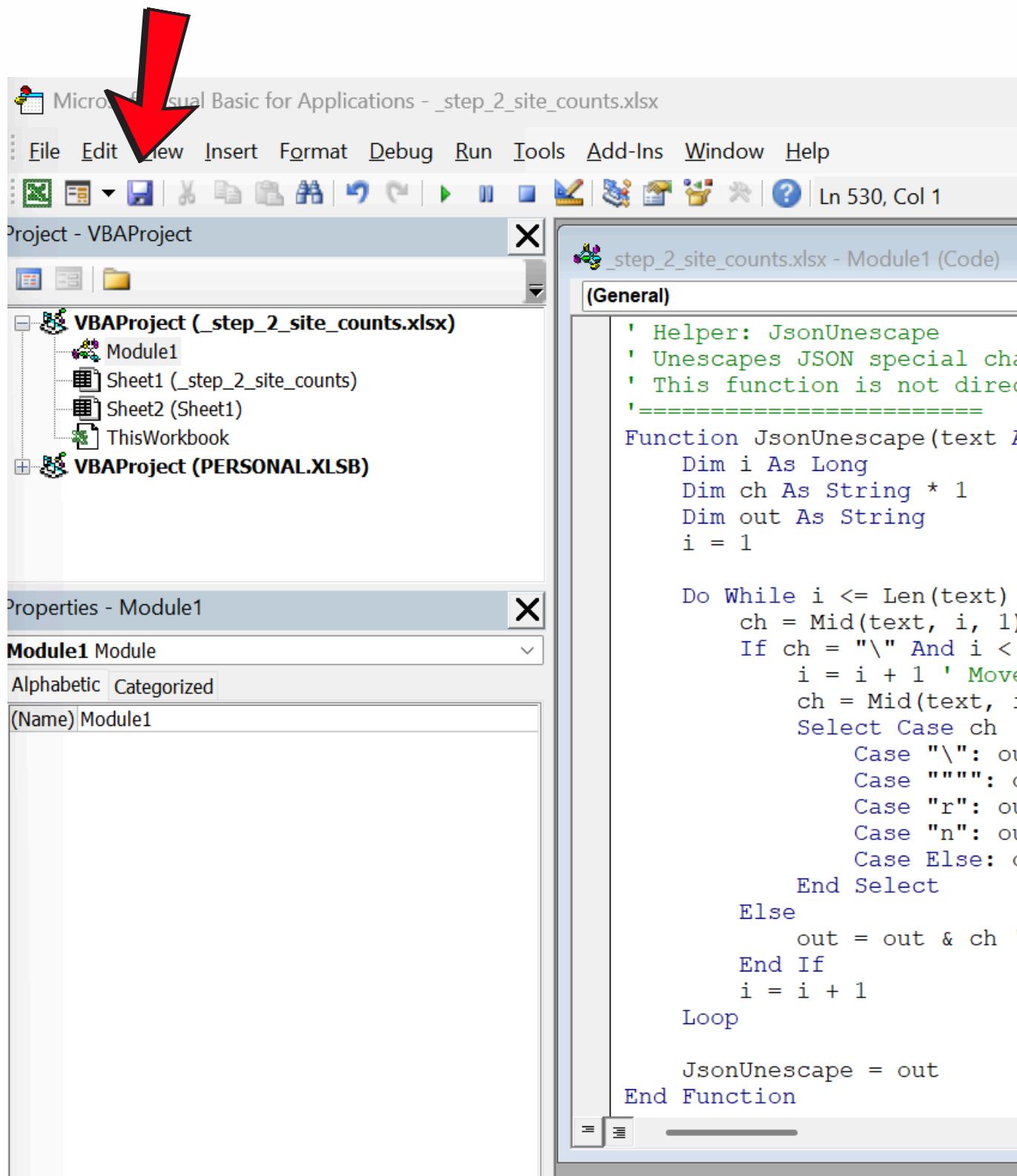
**Please use this file just for processing, and enriching data.**

**If you want to give the data to someone else, move the data into a new file.**

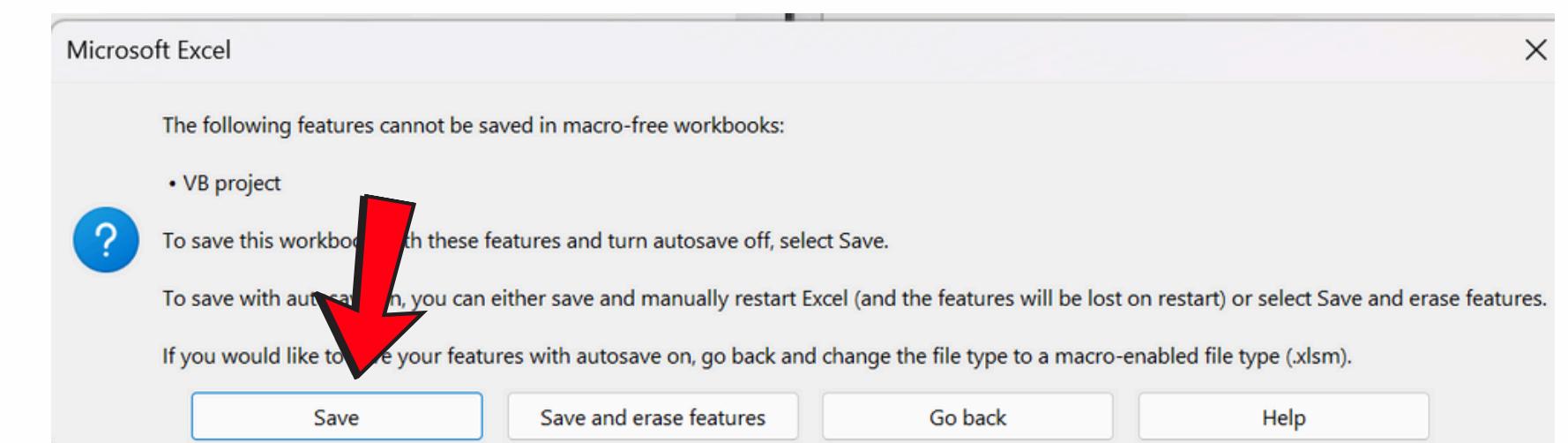
**If the key is compromised you can delete it at any time in openrouter and create a new one**

# Paste the custom function code

## 1. Press Save

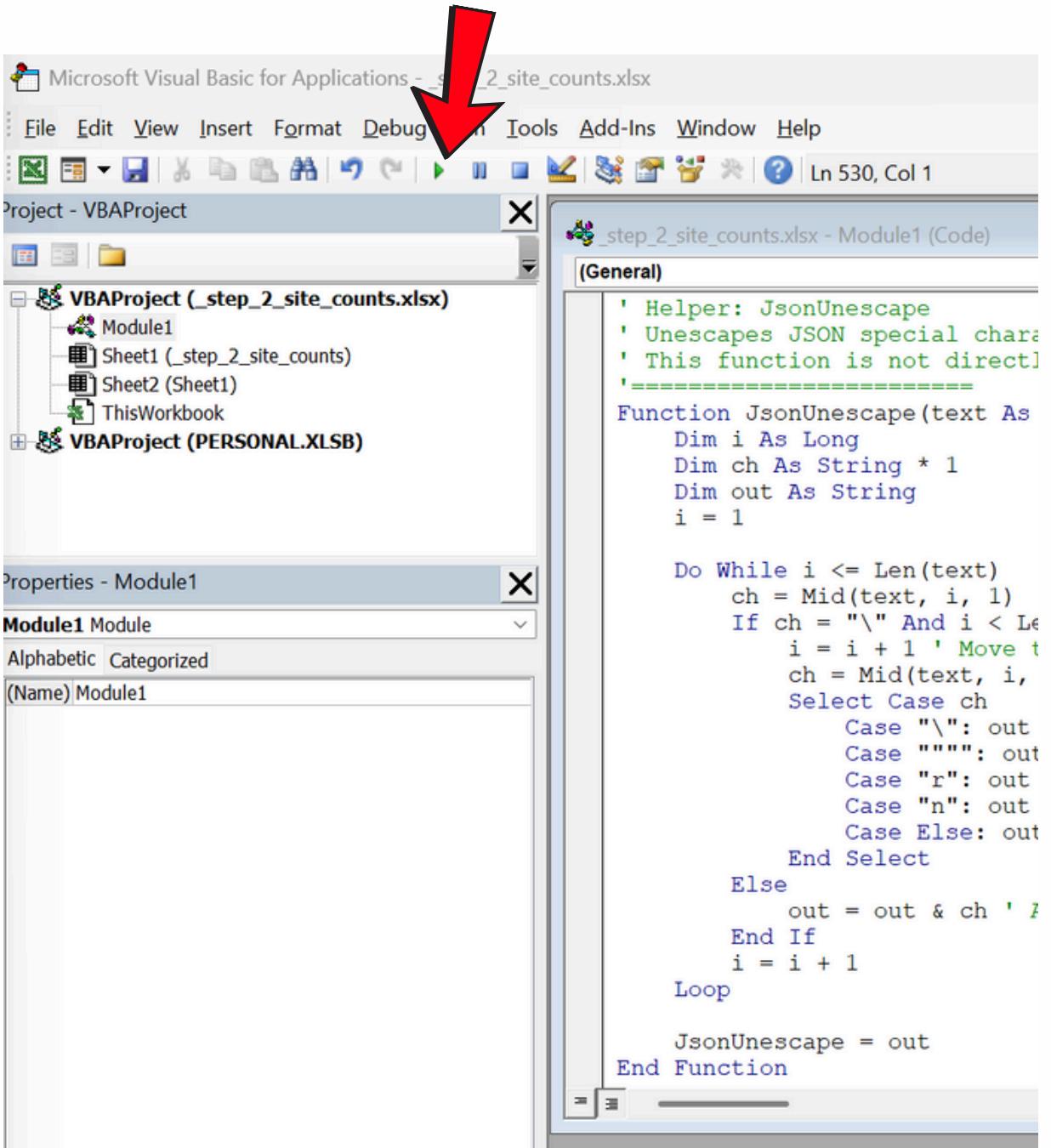


## 2. Save

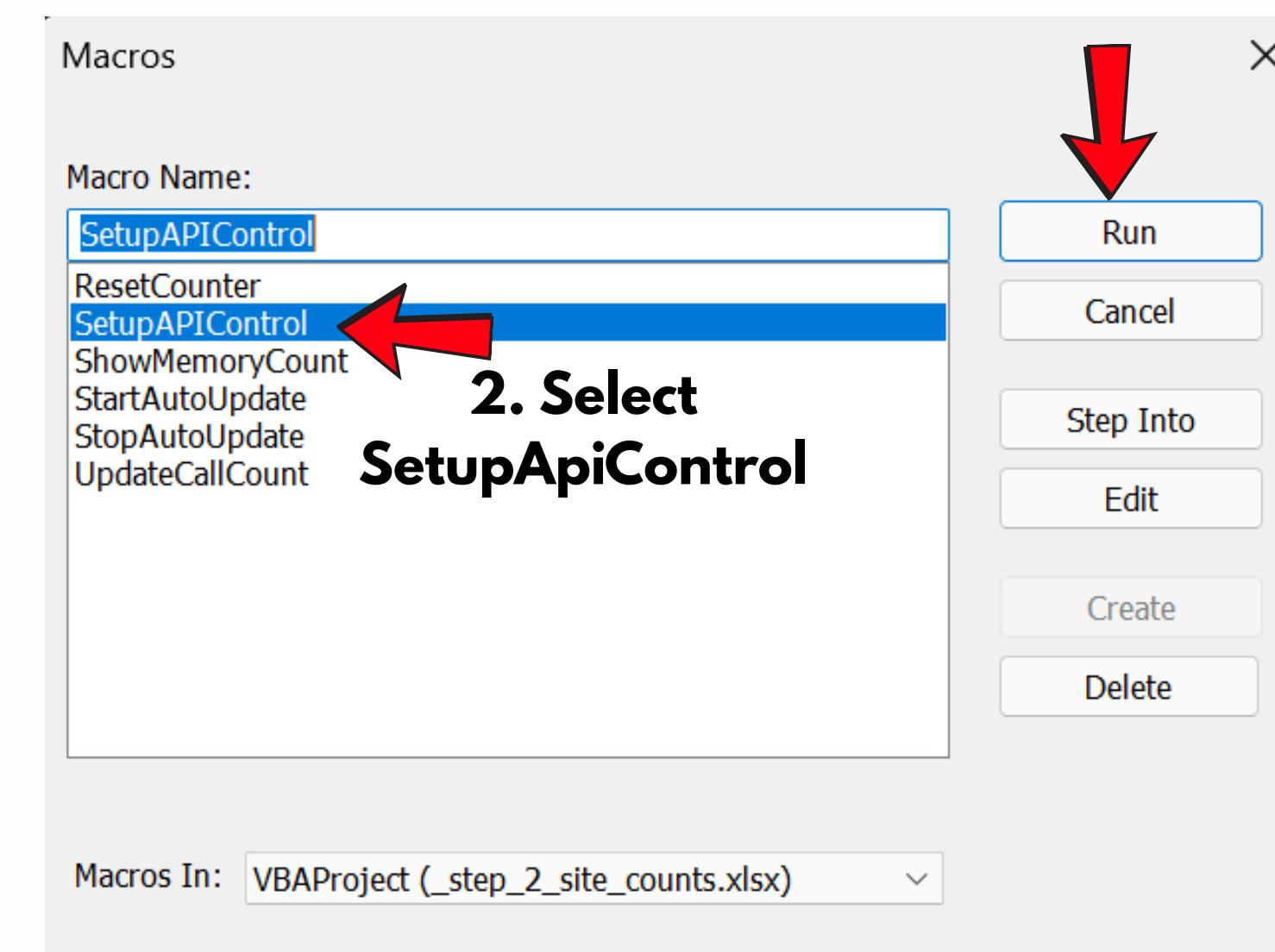


# Paste the custom function code

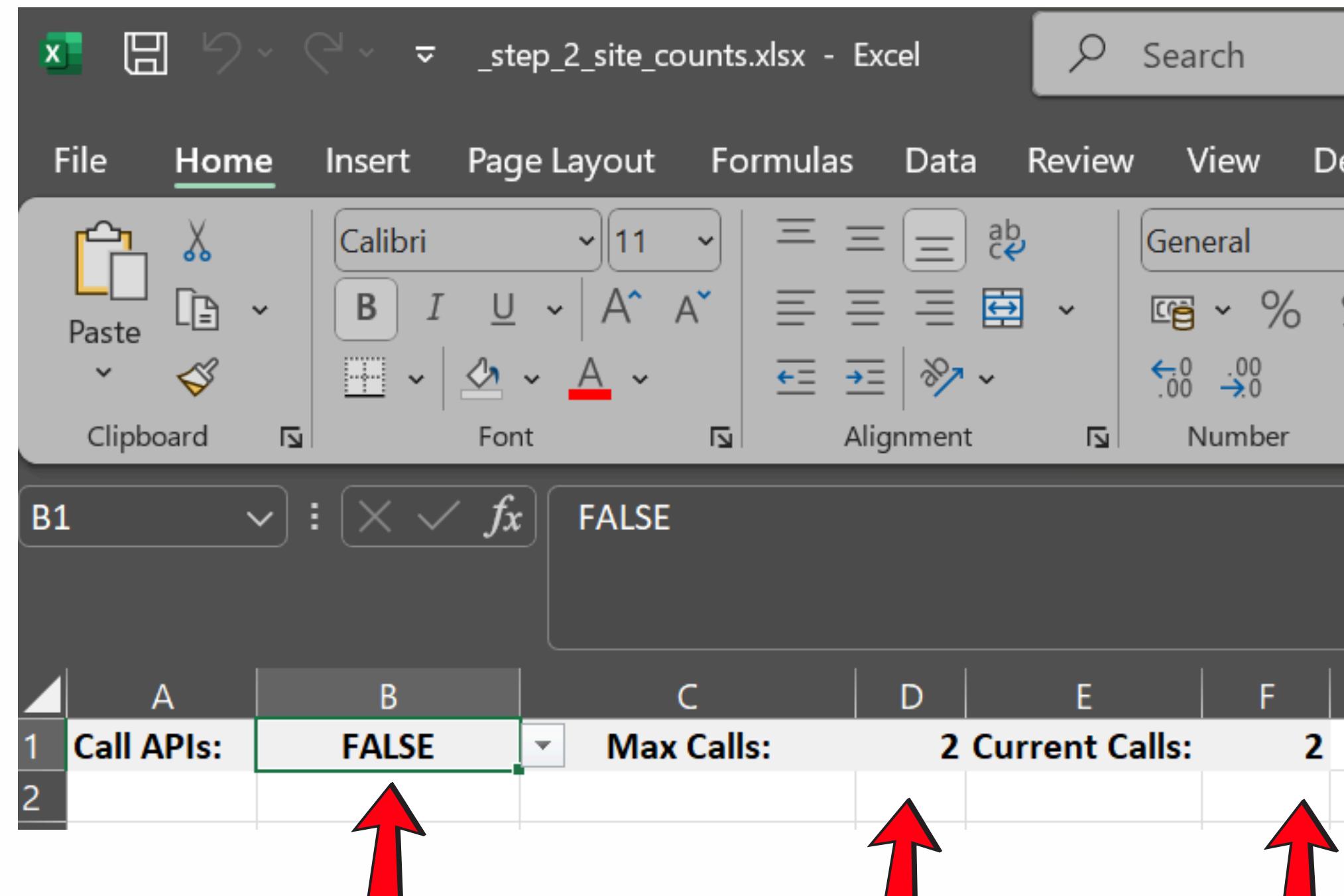
## 1. Press Run



## 3. Run



# LLM API “Cost Control” row will be inserted in your sheet



**Enable and Disable  
API Calls  
(protection)**

**Max LLM API Calls  
(protection)**

**API Calls Count  
since last reset**

# Why the cost control is necessary?

- There are situations when Excel will recalculate all formulas from a table:
  - Insertion of a new row will recalculate rows below
  - Change of formatting
- Also the script has a caching feature for it tries to avoid unnecessary recalculations, but is better to have 2 safety systems than none.

	A	B	C	D	E	F
1	Call APIs:	FALSE		Max Calls:	2 Current Calls:	2
2						

Enable and Disable  
API Calls  
(protection)

Max LLM API Calls  
(protection)

API Calls Count  
since last reset

# Our custom function syntax

**LLM(Prompt; *Temperature*, *Model*)**

Temperature and Model are optional

## Generic Examples:

```
=LLM("Summarize this paragraph: " & A1)  
=LLM("Analyze the trend in cell range A1:A10"; 0,1)  
=LLM("Explain quantum physics simply"; 0,1; "gpt-4o-mini")  
=LLM("Write a poem about data"; 0,9; "claude-3-sonnet")
```

# Template of LLM function for Category Classification

**Do not forget to replace the A2 cell with your input cell**

```
=LLM(  
"create an AI category for the company. Reply only with the category name and nothing else before or after. Adhere strictly to this. "&  
"Assign one category only from this standardized taxonomy: "&  
"- AI_Development_Tools (APIs, SDKs, development platforms) "&  
"- AI_Models_LLMs (base models, foundation models, model providers) "&  
"- AI_Applications_chatbot (chatbots) "&  
"- AI_Applications_voice (voice agents) "&  
"- AI_Applications_image (image generation) "&  
"- AI_Applications_video (video generation) "&  
"- AI_Applications_other (other specific AI tools) "&  
"- AI_Education (courses, tutorials, learning platforms) "&  
"- AI_Research (papers, academic institutions, research organizations) "&  
"- AI_News_Media (industry news, AI journalism, trend analysis) "&  
"- AI_Infrastructure (cloud platforms, MLOps, deployment tools) "&  
"- AI_Community (forums, social platforms, professional networks) "&  
"- Non_AI_Tech (general tech sites with minimal AI focus) "&  
"- Non_AI_General (completely unrelated to AI): "&  
"site=& A2; 0,1; "gpt-4.1-nano")
```

# Template of LLM function for Category Classification

**Temperature Setting: 0 to 0.2**

```
=LLM(  
    "create an AI category for the company. Reply only with the category name and nothing else before or after. Adhere strictly to this. "&  
    "Assign one category only from this standardized taxonomy: "&  
    "- AI_Development_Tools (APIs, SDKs, development platforms) "&  
    "- AI_Models_LLMs (base models, foundation models, model providers) "&  
    "- AI_Applications_chatbot (chatbots) "&  
    "- AI_Applications_voice (voice agents) "&  
    "- AI_Applications_image (image generation) "&  
    "- AI_Applications_video (video generation) "&  
    "- AI_Applications_other (other specific AI tools) "&  
    "- AI_Education (courses, tutorials, learning platforms) "&  
    "- AI_Research (papers, academic institutions, research organizations) "&  
    "- AI_News_Media (industry news, AI journalism, trend analysis) "&  
    "- AI_Infrastructure (cloud platforms, MLOps, deployment tools) "&  
    "- AI_Community (forums, social platforms, professional networks) "&  
    "- Non_AI_Tech (general tech sites with minimal AI focus) "&  
    "- Non_AI_General (completely unrelated to AI): "&  
    "site="& A2; 0; "gpt-4o-mini")
```

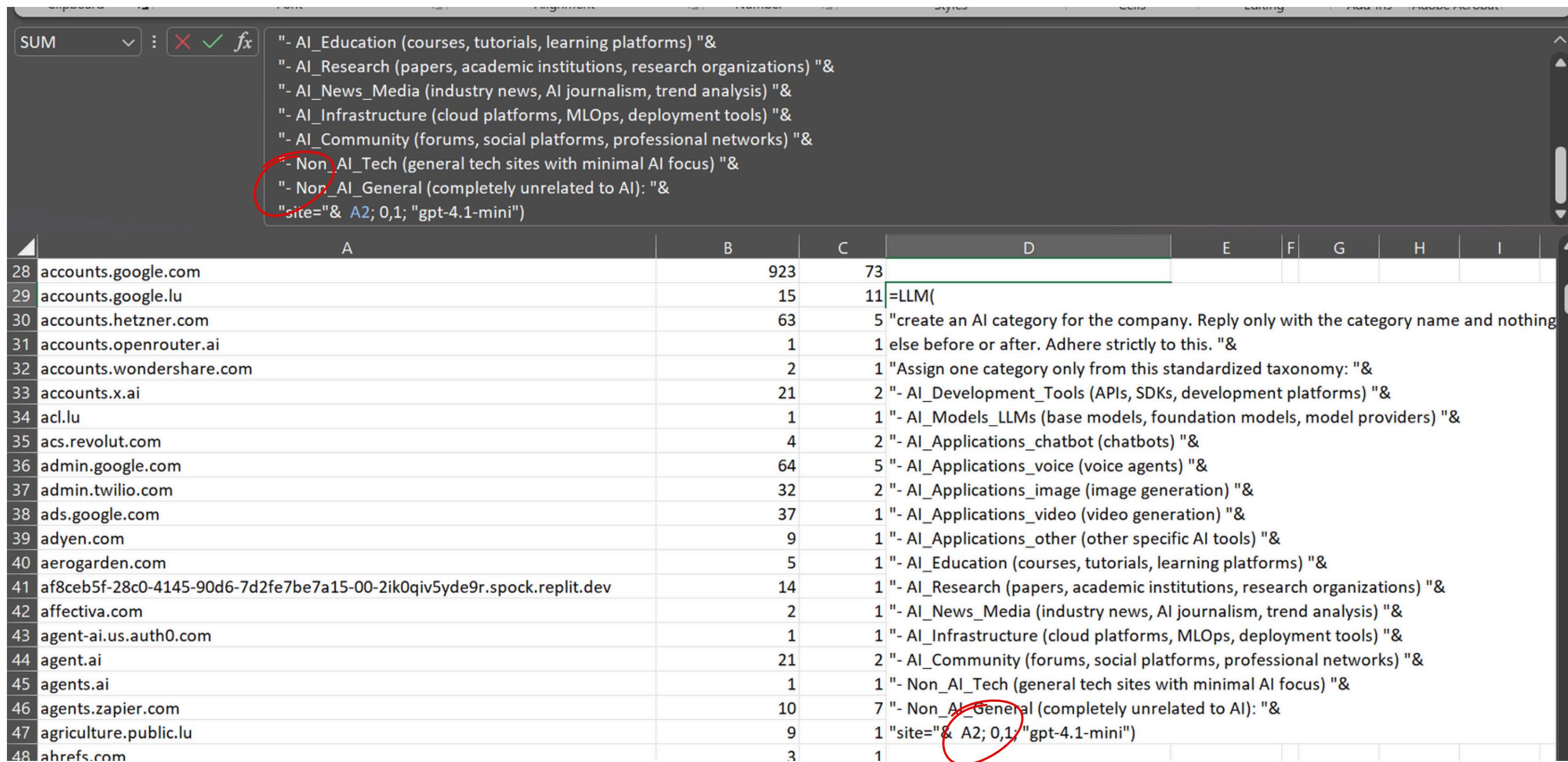
For this case is recommended 0

# Template of LLM function for Description Generation

**Do not forget to replace the A3 cell with your  
input cell**

```
=LLM("Create a Website_Description: "&  
"- Provide a concise 15–25 word description covering: primary purpose, target audience, and core value proposition "&  
"- Format: '[Company/Platform] - [What they do] for [target audience] offering [key services/products]' "&  
"When evaluating, the most relevant is the base domain name. So write the first part of description about it" &  
"Then include the subdomain in the evaluation if adds more clarification in the" &  
" context of the domain name. (example: google.com -> tech company " &  
"and gemini.google.com -> AI chatbot service, Foundational model)" &  
"do not put commas , in the description" &  
"Reply just with the description and nothing else before or after. Website=" & A3; "0.0"; "perplexity/llama-3.1-sonar-small-128k-  
online")
```

- 1. Paste the Function Template into an empty column, empty cell.**
- 2. Edit the “A2” → in my case will be “A29”**
- 3. Press enter after you put the right source cell**



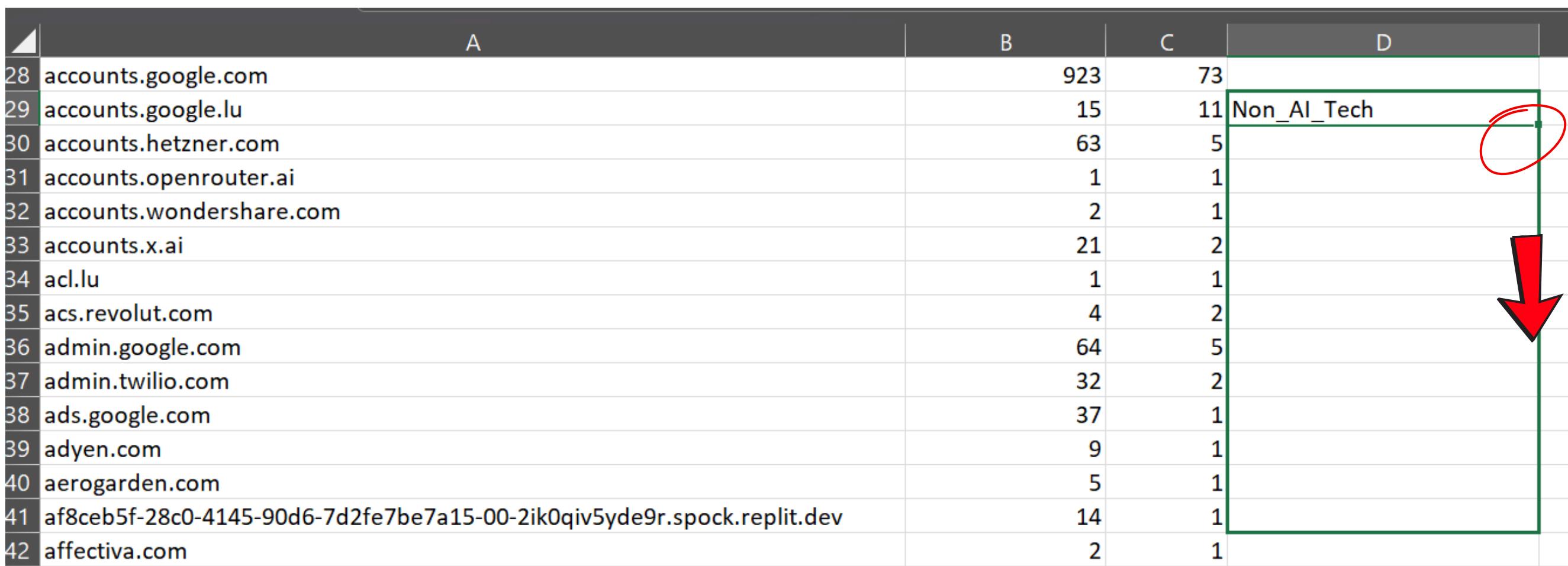
The screenshot shows a Google Sheets interface with a dark theme. A red circle highlights the formula bar at the top, which contains the following function template:

```
-- AI_Education (courses, tutorials, learning platforms) "&
-- AI_Research (papers, academic institutions, research organizations) "&
-- AI_News_Media (industry news, AI journalism, trend analysis) "&
-- AI_Infrastructure (cloud platforms, MLOps, deployment tools) "&
-- AI_Community (forums, social platforms, professional networks) "&
- Non_AI_Tech (general tech sites with minimal AI focus) "&
- Non_AI_General (completely unrelated to AI): "&
"site=& A2; 0,1; "gpt-4.1-mini")
```

A red circle also highlights the cell reference "A2" in the formula, indicating it should be replaced with "A29". The spreadsheet has three columns labeled A, B, and C. Column A lists various website domains, and column C contains the formula applied to each row. The formula in cell D11 is =LLM(, and the result is "create an AI category for the company. Reply only with the category name and nothing else before or after. Adhere strictly to this. "&".

	A	B	C	D	E	F	G	H	I
28	accounts.google.com	923	73	=LLM(					
29	accounts.google.lu	15	11	5 "create an AI category for the company. Reply only with the category name and nothing					
30	accounts.hetzner.com	63	1	else before or after. Adhere strictly to this. "&					
31	accounts.openrouter.ai	1	2	1 "Assign one category only from this standardized taxonomy: "&					
32	accounts.wondershare.com	2	21	2 "- AI_Development_Tools (APIs, SDKs, development platforms) "&					
33	accounts.x.ai	21	1	1 "- AI_Models_LLMs (base models, foundation models, model providers) "&					
34	acl.lu	1	1	2 "- AI_Applications_chatbot (chatbots) "&					
35	acs.revolut.com	4	2	5 "- AI_Applications_voice (voice agents) "&					
36	admin.google.com	64	32	2 "- AI_Applications_image (image generation) "&					
37	admin.twilio.com	32	37	1 "- AI_Applications_video (video generation) "&					
38	ads.google.com	37	9	1 "- AI_Applications_other (other specific AI tools) "&					
39	adyen.com	9	5	1 "- AI_Education (courses, tutorials, learning platforms) "&					
40	aerogarden.com	5	14	1 "- AI_Research (papers, academic institutions, research organizations) "&					
41	af8ceb5f-28c0-4145-90d6-7d2fe7be7a15-00-2ik0qiv5yde9r.spock.replit.dev	14	2	1 "- AI_News_Media (industry news, AI journalism, trend analysis) "&					
42	affectiva.com	2	1	1 "- AI_Infrastructure (cloud platforms, MLOps, deployment tools) "&					
43	agent-ai.us.auth0.com	1	21	2 "- AI_Community (forums, social platforms, professional networks) "&					
44	agent.ai	21	1	1 "- Non_AI_Tech (general tech sites with minimal AI focus) "&					
45	agents.ai	1	10	7 "- Non_AI_General (completely unrelated to AI): "&					
46	agents.zapier.com	10	9	1 "site=& A2; 0,1; "gpt-4.1-mini")					
47	agriculture.public.lu	9	1						
48	ahrefs.com	1							

- 1.The LLM normally filled in the value of that first cell**
- 2.Drag the cell by the corner downwards**
- 3.This will trigger the function to call the LLM for all the cells below**
- 4..... Filling them all might take a while be patient**



	A	B	C	D
28	accounts.google.com	923	73	
29	accounts.google.lu	15	11	Non_AI_Tech
30	accounts.hetzner.com	63	5	
31	accounts.openrouter.ai	1	1	
32	accounts.wondershare.com	2	1	
33	accounts.x.ai	21	2	
34	acl.lu	1	1	
35	acs.revolut.com	4	2	
36	admin.google.com	64	5	
37	admin.twilio.com	32	2	
38	ads.google.com	37	1	
39	adyen.com	9	1	
40	aerogarden.com	5	1	
41	af8ceb5f-28c0-4145-90d6-7d2fe7be7a15-00-2ik0qiv5yde9r.spock.replit.dev	14	1	
42	affectiva.com	2	1	

# ...after a while

	A	B	C	D
28	accounts.google.com	923	73	
29	accounts.google.lu	15	11 Non_AI_General	
30	accounts.hetzner.com	63	5 Non_AI_Tech	
31	accounts.openrouter.ai	1	1 AI_Development_Tools	✗
32	accounts.wondershare.com	2	1 AI_Applications_other	
33	accounts.x.ai	21	2 AI_Applications_other	
34	acl.lu	1	1 AI_Research	✗
35	acs.revolut.com	4	2 AI_Applications_other	✗
36	admin.google.com	64	5 Non_AI_General	
37	admin.twilio.com	32	2 AI_Applications_other	
38	ads.google.com	37	1 Non_AI_General	
39	adyen.com	9	1 Non_AI_Tech	
40	aerogarden.com	5	1 Non_AI_General	
41	af8ceb5f-28c0-4145-90d6-7d2fe7be7a15-00-2ik0qiv5yde9r.spock.replit.dev	14	1 AI_Applications_chatbot	✗
42	affectiva.com	2	1 AI_Applications_other	
43	agent-ai.us.auth0.com	1	1 AI_Applications_chatbot	
44	agent.ai	21	2 AI_Applications_chatbot	✗
45	agents.ai	1	1 AI_Applications_chatbot	
46	agents.zapier.com	10	7 AI_Applications_other	
47	agriculture.public.lu	9	1	
48	ahrefs.com	3	1	
49	ai-service-demos.on-aws.com	2	1	

For this example I used intentionally a local **mistral-small-3.1-24b-instruct-2503**, which is not particularly good at the task of recognizing websites

So categorization is pretty low quality for now.  
But the function LLM works!

# **Models evaluation**

**openai/gpt-4o-mini-search-preview**

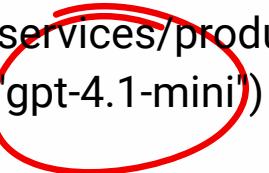
**openai/gpt-4o-mini**

**openai/gpt-4.1-mini**

**openai/gpt-4.1-nano**

**mistralai/mistral-small-3.1-24b-instruct**

**meta-llama/llama-4-maverick**

=LLM("Create a Website\_Description: "&  
"- Provide a concise 15–25 word description covering: primary purpose, target audience, and core value proposition "&  
"- Format: `"[Company/Platform] - [What they do] for [target audience] offering [key services/products]"` "&  
"Reply just with the description and nothing else before or after. Site=" & A3; "0.1"; 'gpt-4.1-mini')  


- 1. Add a variety of the “edge cases” websites**
- 2. Add the selection of models to test**
- 3. Add your template prompt and put as parameters the website COL 1 cells and the models COL B cells**

Call APIs:	FALSE		Max Calls:	
			Category	Description
acl.lu	openai/gpt-4o-mini-search-preview	=LLM(		
acl.lu	openai/gpt-4o-mini	"create an AI category for the company. Reply only with the category name and nothing else before or after. Adhe		
acl.lu	openai/gpt-4.1-mini	"Assign one category only from this standardized taxonomy: "&		
acl.lu	openai/gpt-4.1-nano	"- AI_Development_Tools (APIs, SDKs, development platforms) "&		
acl.lu	mistralai/mistral-small-3.1-24b-instruct	"- AI_Models_LLMs (base models, foundation models, model providers) "&		
acl.lu	meta-llama/llama-4-maverick	"- AI_Applications_chatbot (chatbots) "&		
acs.revolut.com	openai/gpt-4o-mini-search-preview	"- AI_Applications_voice (voice agents) "&		
acs.revolut.com	openai/gpt-4o-mini	"- AI_Applications_image (image generation) "&		
acs.revolut.com	openai/gpt-4.1-mini	"- AI_Applications_video (video generation) "&		
acs.revolut.com	openai/gpt-4.1-nano	"- AI_Applications_other (other specific AI tools) "&		
acs.revolut.com	mistralai/mistral-small-3.1-24b-instruct	"- AI_Education (courses, tutorials, learning platforms) "&		
acs.revolut.com	meta-llama/llama-4-maverick	"- AI_Research (papers, academic institutions, research organizations) "&		
11x.ai	openai/gpt-4o-mini-search-preview	"- AI_News_Media (industry news, AI journalism, trend analysis) "&		
11x.ai	openai/gpt-4o-mini	"- AI_Infrastructure (cloud platforms, MLOps, deployment tools) "&		
11x.ai	openai/gpt-4.1-mini	"- AI_Community (forums, social platforms, professional networks) "&		
11x.ai	openai/gpt-4.1-nano	"- Non_AI_Tech (general tech sites with minimal AI focus) "&		
11x.ai	mistralai/mistral-small-3.1-24b-instruct	"- Non_AI_General (completely unrelated to AI): "&		
11x.ai	meta-llama/llama-4-maverick	"site=& A4; 0,1; B4)		
ainnovate.tech	openai/gpt-4o-mini-search-preview			
ainnovate.tech	openai/gpt-4o-mini			
ainnovate.tech	openai/gpt-4.1-mini			
ainnovate.tech	openai/gpt-4.1-nano			
ainnovate.tech	mistralai/mistral-small-3.1-24b-instruct			
ainnovate.tech	meta-llama/llama-4-maverick			
aistudio.google.com	openai/gpt-4o-mini-search-preview			
aistudio.google.com	openai/gpt-4o-mini			
aistudio.google.com	openai/gpt-4.1-mini			

A	B	C	D
Call APIs:	TRUE	Max Calls:	100
acl.lu	openai/gpt-4o-mini-search-preview	Category	Description
acl.lu	openai/gpt-4o-mini	Non_AI_Tech	=LLM("Create a Website_Description: "&"- Provide a concise 15–25 word description covering: primary purpose, target audience, and core value proposition "&"- Format: '[Company/Platform] - [What they do] for [target audience] offering [key services/products]' "&"Reply just with the description and nothing else before or after. Site=" & A4; "0.0"; B4)
acl.lu	openai/gpt-4.1-mini	AI_Research	
acl.lu	openai/gpt-4.1-nano	AI_Research	
acl.lu	mistralai/mistral-small-3.1-24b-instruct	Non_AI_Tech	
acl.lu	meta-llama/llama-4-maverick	AI_Research	
acs.revolut.com	openai/gpt-4o-mini-search-preview	Non_AI_Tech	
acs.revolut.com	openai/gpt-4o-mini	Non_AI_General	
acs.revolut.com	openai/gpt-4.1-mini	AI_Applications_other	
acs.revolut.com	openai/gpt-4.1-nano	Non_AI_Tech	
acs.revolut.com	mistralai/mistral-small-3.1-24b-instruct	AI_Applications_other	
acs.revolut.com	meta-llama/llama-4-maverick	AI_Applications_chatbot	
11x.ai	openai/gpt-4o-mini-search-preview	AI_Applications_other	
11x.ai	openai/gpt-4o-mini	AI_Applications_chatbot	
11x.ai	openai/gpt-4.1-mini	AI_Applications_chatbot	
11x.ai	openai/gpt-4.1-nano	AI_Models_LLMs	
11x.ai	mistralai/mistral-small-3.1-24b-instruct	AI_Applications_other	
11x.ai	meta-llama/llama-4-maverick	AI_Development_Tools	
ainnovate.tech	openai/gpt-4o-mini-search-preview	AI_Applications_other	
ainnovate.tech	openai/gpt-4o-mini	AI_Applications_other	
ainnovate.tech	openai/gpt-4.1-mini	AI_Applications_other	
ainnovate.tech	openai/gpt-4.1-nano	AI_Development_Tools	
ainnovate.tech	mistralai/mistral-small-3.1-24b-instruct	AI_Applications_other	
ainnovate.tech	meta-llama/llama-4-maverick	AI_Development_Tools	

A	B	C	D
Call APIs:	TRUE	Max Calls:	100
acl.lu	openai/gpt-4o-mini-search-preview	Category	Description
acl.lu	openai/gpt-4o-mini	Non_AI_Tech	✓ Automobile Club du Luxembourg (ACL) - Provides comprehensive mobility services for members, including roadside assistance, vehicle diagnostics, and travel planning.
acl.lu	openai/gpt-4.1-mini	AI_Research	✗ ACL - Provides legal and compliance solutions for businesses offering expert advice, training, and resources to navigate regulatory challenges.
acl.lu	openai/gpt-4.1-nano	AI_Research	✗ ACL.lu - Provides legal and administrative support for Luxembourg residents offering expert advice, resources, and personalized assistance.
acl.lu	openai/gpt-4.1-nano	Non_AI_Tech	✗ ACL.lu - Luxembourg's legal information platform for professionals offering comprehensive legal resources, updates, and tools to support legal practice.
acl.lu	mistralai/mistral-small-3.1-24b-instruct	AI_Research	✗ ACL.lu - Empowers legal professionals in Luxembourg with comprehensive legal resources and tools, offering up-to-date legislation, case law, and expert analysis.
acl.lu	meta-llama/llama-4-maverick	AI_Research	✗ ACL - Providing Luxembourg's public administration with IT services and solutions, offering secure, innovative, and integrated digital infrastructure and support.
acs.revolut.com	openai/gpt-4o-mini-search-preview	Non_AI_Tech	
acs.revolut.com	openai/gpt-4o-mini	Non_AI_General	

We add also the description enrichment to see more details about how each LLM sees the website

# I added also some of the bigger models

## For many use cases a small model with tools > big model

acl.lu	openai/gpt-4.1		<span style="color: red;">X</span>	acl.lu - Provides IT security solutions for businesses and organizations, offering cybersecurity consulting, penetration testing, and compliance services.
acl.lu	deepseek/deepseek-prover-v2		<span style="color: red;">X</span>	'ACL.lu - A Luxembourg-based company offering comprehensive IT solutions and services for businesses, focusing on digital transformation and cybersecurity.'
acl.lu	deepseek/deepseek-r1-0528		<span style="color: green;">✓</span>	ACL - Luxembourg's premier automobile club for motorists, offering roadside assistance, insurance, travel planning, and road safety advocacy services.
acl.lu	anthropic/clause-sonnet-4		<span style="color: green;">✓</span>	ACL - Automotive Club of Luxembourg providing roadside assistance, insurance, and travel services for Luxembourg drivers and travelers.
acl.lu	google/gemini-2.5-pro-preview		<span style="color: green;">✓</span>	ACL - Provides mobility and travel solutions for drivers and travelers in Luxembourg, offering 24/7 roadside assistance, travel services, and member benefits.
acl.lu	google/gemini-2.5-flash-preview-05-20		<span style="color: red;">X</span>	ACL.lu - Luxembourg's leading independent IT services provider for businesses, offering comprehensive IT solutions and expert consulting.
acl.lu	openai/gpt-4o		<span style="color: red;">X</span>	ACL - IT solutions for businesses offering cybersecurity, cloud services, and digital transformation expertise.
acl.lu	openai/gpt-4o-mini-search-preview	Non_AI_Tech	<span style="color: green;">✓</span>	Automobile Club du Luxembourg (ACL) - Provides comprehensive mobility services for individuals, offering roadside assistance, vehicle diagnostics, rentals, and travel planning.
acl.lu	openai/gpt-4o-mini	AI_Research	<span style="color: red;">X</span>	ACL - Provides legal and compliance solutions for businesses offering expert advice, training, and resources to navigate regulatory challenges.
acl.lu	openai/gpt-4.1-mini	AI_Research	<span style="color: red;">X</span>	ACL.lu - Provides legal and consumer protection information for Luxembourg residents offering guidance, resources, and support services.
acl.lu	openai/gpt-4.1-nano	Non_AI_Tech	<span style="color: red;">X</span>	ACL.lu - Luxembourg's legal platform providing resources and tools for legal professionals and students offering comprehensive legal information and services
acl.lu	mistralai/mistral-small-3.1-24b-instruct	AI_Research	<span style="color: red;">X</span>	ACL.lu - Empowers legal professionals in Luxembourg with comprehensive legal resources and tools, offering up-to-date legislation, case law, and expert analysis.
acl.lu	meta-llama/llama-4-maverick	AI_Community	<span style="color: red;">X</span>	ACL - Providing Luxembourg's public administration with IT services and solutions, offering digital transformation, cybersecurity, and data management expertise.

# Another edge case test:

57639728-2f27-4b8f-af31-f7d6b17e5385-00-  
2zlkfopcbgt0w.janeway.replit.dev

perplexity/llama-3.1-sonar-small-128k-online	AI_Development_Tools	Janeway - A Node.js console REPL for developers, offering an uncaughtException handler and enhanced debugging capabilities.	
openai/gpt-4o-mini-search-preview	Non_AI_Tech	janeway.replit.dev - Node.js console REPL with object inspection for developers offering interactive command line, autocomplete, and clickable	
perplexity/sonar	AI_Development_Tools	Janeway.replit.dev - Develops AI-assisted coding projects for developers offering automated environment setup and deployment services.	
openai/gpt-4.1-nano	AI_Development_Tools	Janeway.replit.dev - An educational platform for developers offering interactive coding tutorials and resources to enhance programming skills.	
mistralai/mistral-small-3.1-24b-instruct	AI_Applications_chatbot	Janeway - Provides a collaborative coding environment for developers, offering real-time code editing and project management.	
qwen/qwen3-235b-a22b	AI_Development_Tools	Replit - Online IDE for developers offering real-time code collaboration, cloud-based environments, and project sharing.	
deepseek/deepseek-r1-0528	AI_Development_Tools	Replit - A deployed web application on Replit's platform for developers and learners, offering cloud-based coding, collaboration, and hosting services.	
google/gemini-2.5-flash-preview-05-20	AI_Applications_other	Janeway Replit - A development platform for coders, offering collaborative coding environments and project hosting.	

# The extended test show even bigger models failing

1	perplexity/llama-3.1-sonar-small-128k-online	\$0.20 / \$0.20	\$0.200	✓
2	mistralai/mistral-small-3.1-24b-instruct	\$0.10 / \$0.30	\$0.200	✗
3	openai/gpt-4.1-nano	\$0.10 / \$0.40	\$0.250	✗
4	qwen/qwen3-235b-a22b	\$0.13 / \$0.60	\$0.365	✗
5	google/gemini-2.5-flash-preview-05-20	\$0.15 / \$0.60	\$0.375	✗
6	openai/gpt-4o-mini-search-preview	\$0.15 / \$0.60	\$0.375	✓
7	openai/gpt-4o-mini	\$0.15 / \$0.60	\$0.375	✗
8	meta-llama/llama-4-maverick	\$0.15 / \$0.60	\$0.375	✗
9	perplexity/sonar	\$1.00 / \$1.00	\$1.000	✓
10	openai/gpt-4.1-mini	\$0.40 / \$1.60	\$1.000	✗
11	mistralai/mistral-medium-3	\$0.40 / \$2.00	\$1.200	✗
12	deepseek/deepseek-r1-0528	\$0.50 / \$2.15	\$1.325	✓
13	deepseek/deepseek-prover-v2	\$0.50 / \$2.18	\$1.340	✗
14	openai/o4-mini	\$1.10 / \$4.40	\$2.750	✓
15	perplexity/sonar-reasoning	\$1.00 / \$5.00	\$3.000	✓
16	mistralai/magistral-medium-2506	\$2.00 / \$5.00	\$3.500	✓
17	openai/gpt-4.1	\$2.00 / \$8.00	\$5.000	✗
18	perplexity/sonar-reasoning-pro	\$2.00 / \$8.00	\$5.000	✓
19	openai/gpt-4o	\$2.50 / \$7.50	\$5.000	✗
20	google/gemini-2.5-pro-preview	\$1.25 / \$10.00	\$5.625	✓
21	anthropic/clause-sonnet-4	\$3.00 / \$15.00	\$9.000	✓

Model	In \$/M	Out \$/M	Blended \$/M	Special Traits
perplexity/llama-3.1-sonar-small-128k-online	0.20	0.20	0.20	128K context, real-time <b>web-search grounding</b>
mistralai/mistral-small-3.1-24B-instruct	0.10	0.30	0.20	<b>Low-latency</b> 24B instruct model, native function-calling
openai/gpt-4.1-nano	0.10	0.40	0.25	<b>Fastest, cheapest</b> GPT-4.1 for low-latency tasks
qwen/qwen3-235B-A22B	0.13	0.60	0.365	235B MoE; " <b>thinking</b> " mode for deep reasoning, 100+ languages
google/gemini-2.5-flash-preview-05-20	0.15	0.60	0.375	<b>Ultra-low-latency</b> "Flash" multimodal preview
openai/gpt-4o-mini-search-preview	0.15	0.60	0.375	Mini GPT-4o with built-in <b>web-search execution</b>
openai/gpt-4o-mini	0.15	0.60	0.375	<b>Cost-efficient</b> mini GPT-4o; multimodal & function-calling
meta-llama/llama-4-maverick	0.15	0.60	0.375	17B MoE <b>generalist</b> ; chat + reasoning + vision + code
perplexity/sonar	1.00	1.00	1.00	<b>Retrieval-grounded</b> Q&A with live citations
openai/gpt-4.1-mini	0.40	1.60	1.00	Compact GPT-4.1; <b>faster/cheaper</b> than full model
mistralai/mistral-medium-3	0.40	2.00	1.20	<b>Enterprise-ready</b> ; SOTA at 8x lower cost
deepseek/deepseek-r1-0528	0.50	2.15	1.325	<b>Open-source R1</b> reasoning model comparable to OpenAI o1
deepseek/deepseek-prover-v2	0.50	2.18	1.340	Formal <b>theorem-proving specialist</b> (Lean 4)
openai/o4-mini	1.10	4.40	2.75	<b>Math/code reasoning</b> ; solved Tier-4 problems at Berkeley
perplexity/sonar-reasoning	1.00	5.00	3.00	<b>Chain-of-thought reasoning</b> with live search
mistralai/magistral-medium-2506	2.00	5.00	3.50	MoE " <b>Magistral</b> " reasoning, CoT, 128K context
openai/gpt-4.1	2.00	8.00	5.00	Full GPT-4.1 for <b>deep reasoning</b> , larger context
perplexity/sonar-reasoning-pro	2.00	8.00	5.00	<b>Premium CoT reasoning</b> with enhanced retrieval
openai/gpt-4o	2.50	7.50	5.00	Flagship <b>multimodal</b> (text-vision-audio) model
google/gemini-2.5-pro-preview	1.25	10.00	5.625	<b>Long-context</b> , advanced reasoning & coding
anthropic/clause-sonnet-4	3.00	15.00	9.00	<b>Creative writing</b> , large context, "Constitutional AI" safety focus
openai/gpt-4.5-preview	75.00	150.00	112.50	<b>Vast knowledge</b> , emotional nuance & creativity (research preview)

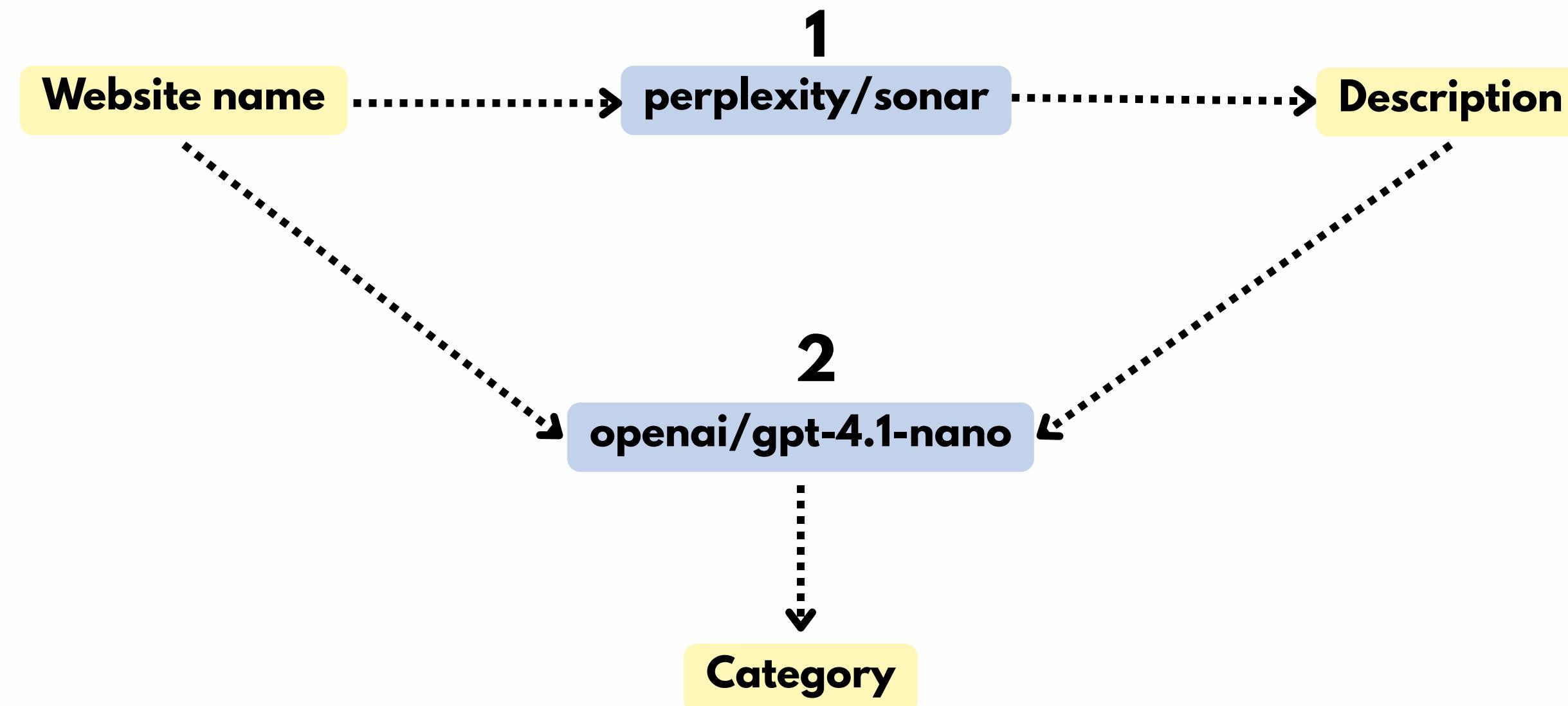
# Project cost - testing 100 rows

<b>openai/gpt-4.1-mini</b>	<b>Description for 100 rows</b>	<b>Tokens Cost \$0.057</b>	<b>Tokens 45k</b>	5x-10x more output (20-40 tokens)
<b>openai/gpt-4o-mini</b>	<b>Category for 100 rows</b>	<b>Tokens Cost \$0.004</b>	<b>Tokens 24k</b>	Very small amount of output tokens: 4-5
<b>perplexity/sonar</b>	<b>Description for 100 rows</b>	<b>Tokens + Req. Cost \$0.517</b>	<b>Tokens 18k</b>	<b>Requests 100 (\$0.5)</b>
<b>deepseek/deepseek-r1-0528</b>	<b>Description for 100 rows</b>	<b>Tokens Cost \$0.360</b>	<b>Tokens 150k</b>	1000-2000 output tokens

# 700 ROWS COST ESTIMATIONS

	MODEL	DESC. COST (\$)	CAT. COST (\$)	SEARCH COST (\$)	GRAND TOTAL COST
mistralai/mistral-small-3.1-24B-instruct		–	0.018	–	<b>0.018</b>
openai/gpt-4.1-nano	⭐	–	0.019	–	<b>0.019</b>
qwen/qwen3-235B-A22B		–	0.025	–	<b>0.025</b>
google/gemini-2.5-flash-preview-05-20		–	0.028	–	<b>0.028</b>
openai/gpt-4o-mini		–	0.028	–	<b>0.028</b>
meta-llama/llama-4-maverick		–	0.028	–	<b>0.028</b>
openai/gpt-4.1-mini		–	0.075	–	<b>0.075</b>
mistralai/mistral-medium-3		–	0.077	–	<b>0.077</b>
deepseek/deepseek-prover-v2		–	0.094	–	<b>0.094</b>
openai/gpt-4.1		–	0.373	–	<b>0.373</b>
openai/gpt-4o		–	0.451	–	<b>0.451</b>
deepseek/deepseek-R1-0528	✓	2.520	0.094	–	<b>2.614</b>
perplexity/llama-3.1-sonar-small-128k-online	✓	0.063	0.034	3.50	<b>3.597</b>
perplexity/sonar	✓ ⭐	0.315	0.168	3.50	<b>3.983</b>
anthropic/clause-sonnet-4	✓	3.675	0.579	–	<b>4.254</b>
openai/o4-mini	✓	5.192	0.205	–	<b>5.397</b>
mistralai/magistral-medium-2506	✓	6.214	0.355	–	<b>6.569</b>
perplexity/sonar-reasoning	✓	5.798	0.193	3.50	<b>9.491</b>
google/gemini-2.5-pro-preview (reasoning)	✓	11.283	0.264	–	<b>11.547</b>
perplexity/sonar-reasoning-pro	✓	9.442	0.373	4.20	<b>14.015</b>
openai/gpt-4o-mini-search-preview	✓	0.150	0.028	17.50	<b>17.678</b>
gpt-4.5-preview	✓	40.688	13.066	–	<b>53.754</b>

# We choose the final solution: LLM to use and processing flow for enrichment



# Final enrichment PROMPT for category

**We will use both website column and description column for classification**

```
=LLM(  
"create an AI category for the company. Reply only with the category name and nothing else before or after.  
Adhere strictly to this."&  
"Assign one category only from this standardized taxonomy:"&  
"- AI_Development_Tools (APIs, SDKs, development platforms)"&  
"- AI_Models_LLMs (base models, foundation models, model providers)"&  
"- AI_Applications_chatbot (chatbots)"&  
"- AI_Applications_voice (voice agents)"&  
"- AI_Applications_image (image generation)"&  
"- AI_Applications_video (video generation)"&  
"- AI_Applications_other (other specific AI tools)"&  
"- AI_Education (courses, tutorials, learning platforms)"&  
"- AI_Research (papers, academic institutions, research organizations)"&  
"- AI_News_Media (industry news, AI journalism, trend analysis)"&  
"- AI_Infrastructure (cloud platforms, MLOps, deployment tools)"&  
"- AI_Community (forums, social platforms, professional networks)"&  
"- Non_AI_Tech (general tech sites with minimal AI focus)"&  
"- Non_AI_General (completely unrelated to AI): "&  
"site=""& A299&" description="" & E2998"""; 0,1; "openai/gpt-4.1-nano")
```

# Creating dashboards and reports

## PROMPT

# TASK: Website Interaction Analysis & AI Ecosystem Visualization

### ## OBJECTIVE

Analyze this CSV file that contains my browser history over the last 3 months, each site having a category attribute. Please identify key AI-related resources, understand my research focus areas, my working tools and habits and visualize engagement patterns that reveal insights about my AI learning journey and professional interests.

### How to work:

- Use all the data given.
- First check programmatically all the existing categories to understand the content
- Count programmatically the rows
- After you know the possible categories, make a plan for the charts how to extract the data for them
- Create the charts programmatically for all the data

### ### Analysis Charts

- Top 15 most-visited AI sites (bar chart) by:
  - total interactions
  - total days
- Engagement intensity heatmap (interactions vs. distinct days) to identify deep-dive vs. casual browsing patterns
- Site category distribution (pie/donut chart) showing AI subdomain focus areas (by days)
- Engagement pattern analysis (high frequency + high day count = core resources)
- Time investment distribution across AI categories
- Discovery vs. reference pattern identification (low days/high interactions vs. high days/varied interactions)
- For the AI categories that are used a lot (top 5), list also the websites in the category in a separate chart
- Add any other chart if you find it useful or insightful given the data we have

### ### Site Classification & Categorization

- Use existing site categorization but also create overarching categories for high level analyses
- Identify AI focus areas

### ## TECHNICAL REQUIREMENTS

- Use modern, professional color schemes (avoid default chart colors)
- Implement consistent typography and spacing
- Include data labels and clear legends

### ## INPUT DATA

Attached below browser history statistics

### ## OUTPUT FORMAT

- Brief executive summary (2–3 sentences) of key findings
- Individual chart sections with contextual analysis
- Output directly everything in one shot do not ask me anything

# Claude

## Executive Summary

298

AI Sites Visited

13,607

Total Interactions

1,078

Browsing Days

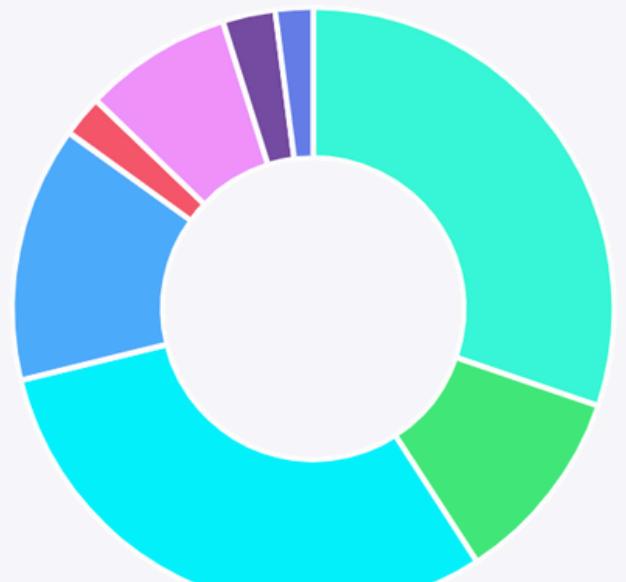
12.6

Avg. Interactions/Day

**Key Finding:** Your AI engagement patterns reveal a sophisticated ecosystem focused on development tools (34% of interactions), community engagement through LinkedIn, and practical AI applications. You demonstrate both deep-dive research behavior and consistent daily usage across multiple AI platforms, indicating a professional AI practitioner profile.

## AI Category Distribution

Focus areas by total browsing days

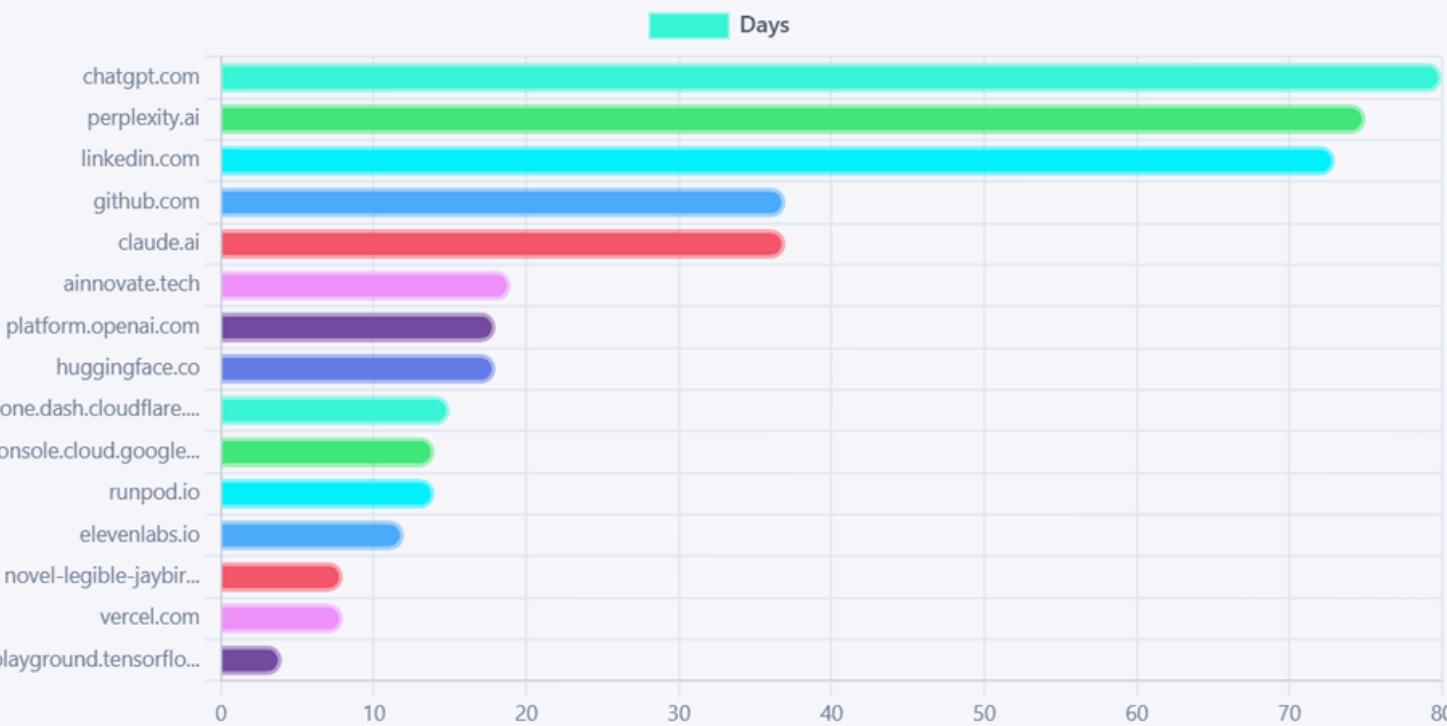


Legend for AI Category Distribution:

- AI Dev Tools
- AI Community
- AI Chatbots
- AI Infrastructure
- AI Voice Tools
- AI Applications
- AI Video Tools
- AI Image Tools

## Top 15 AI Sites by Days

Consistency of engagement pattern

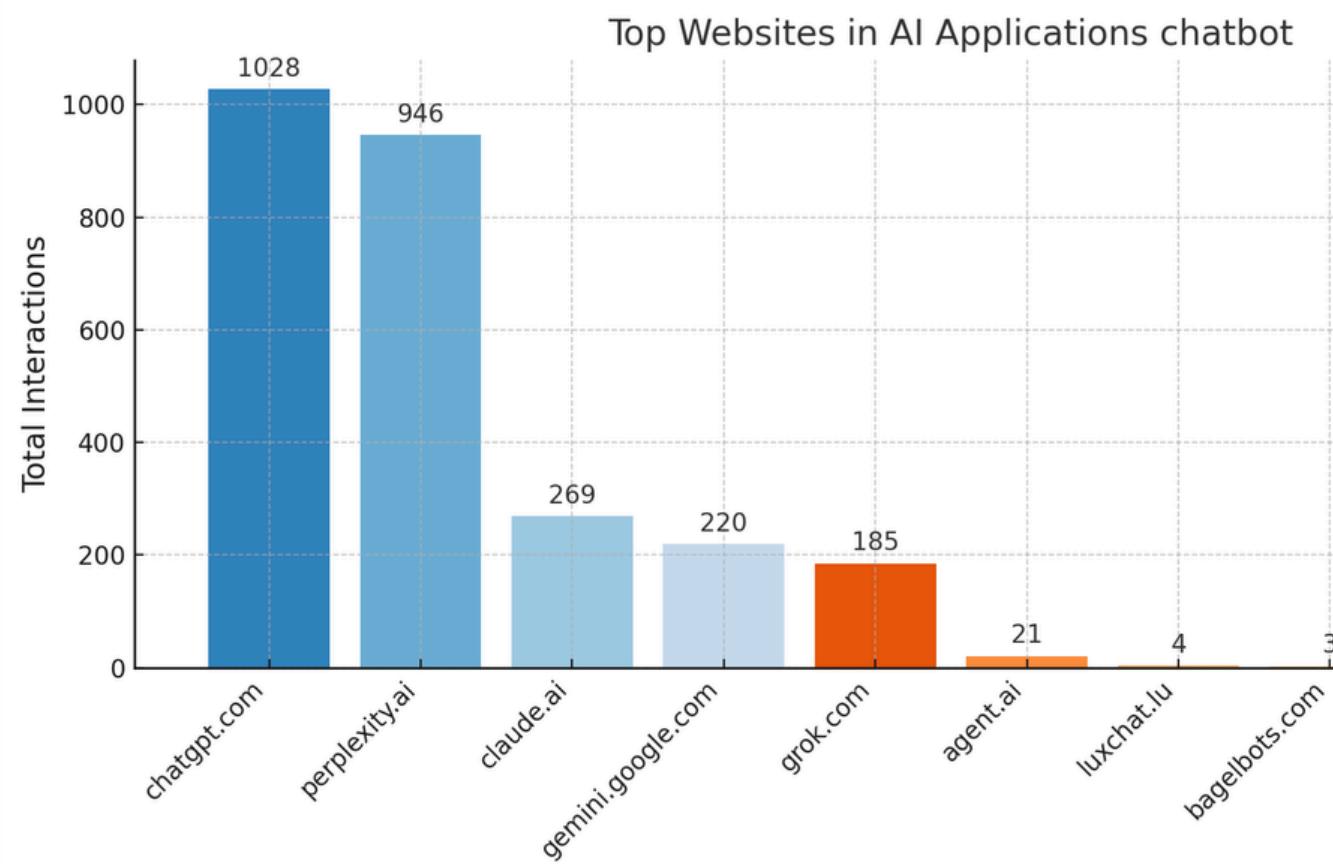
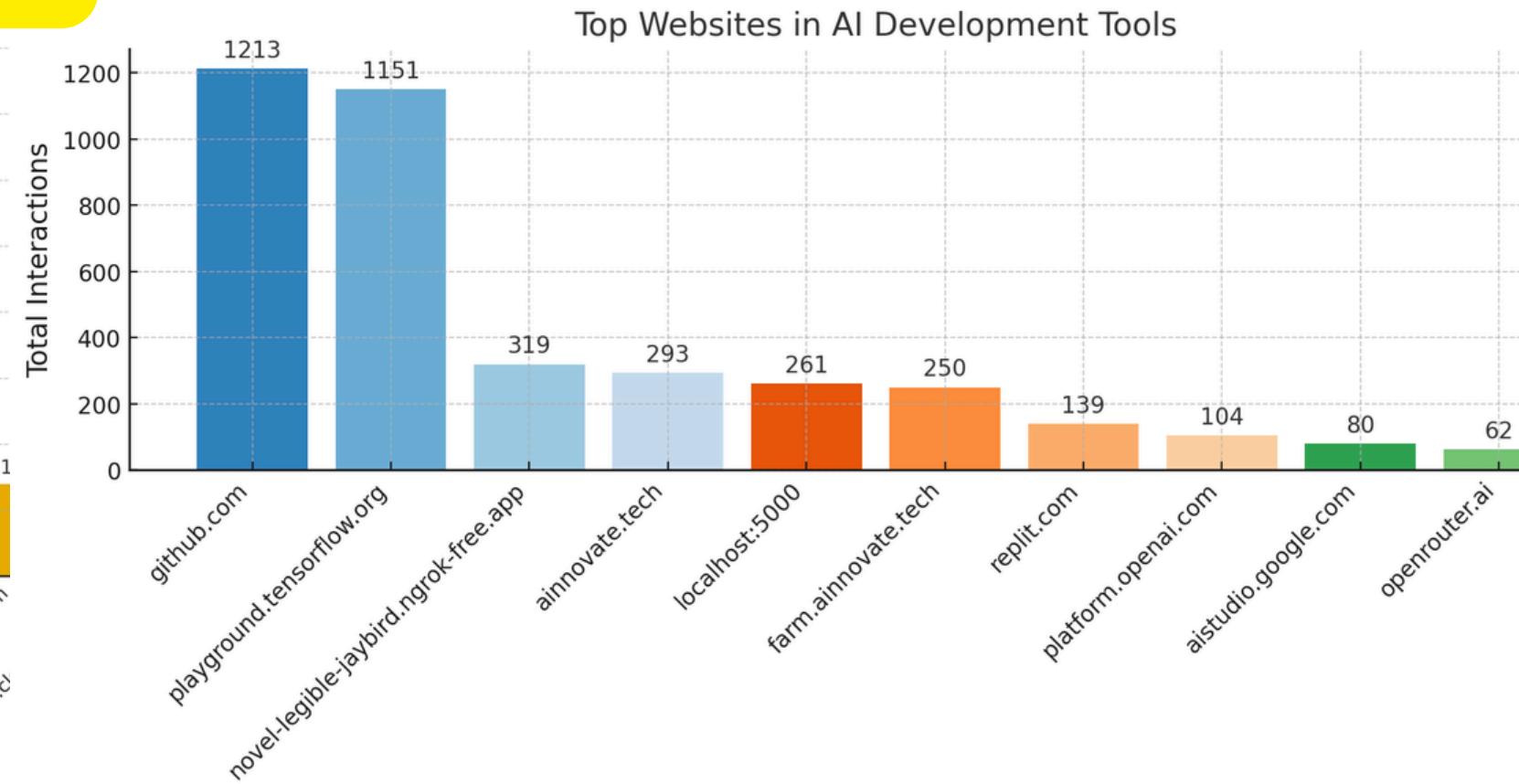
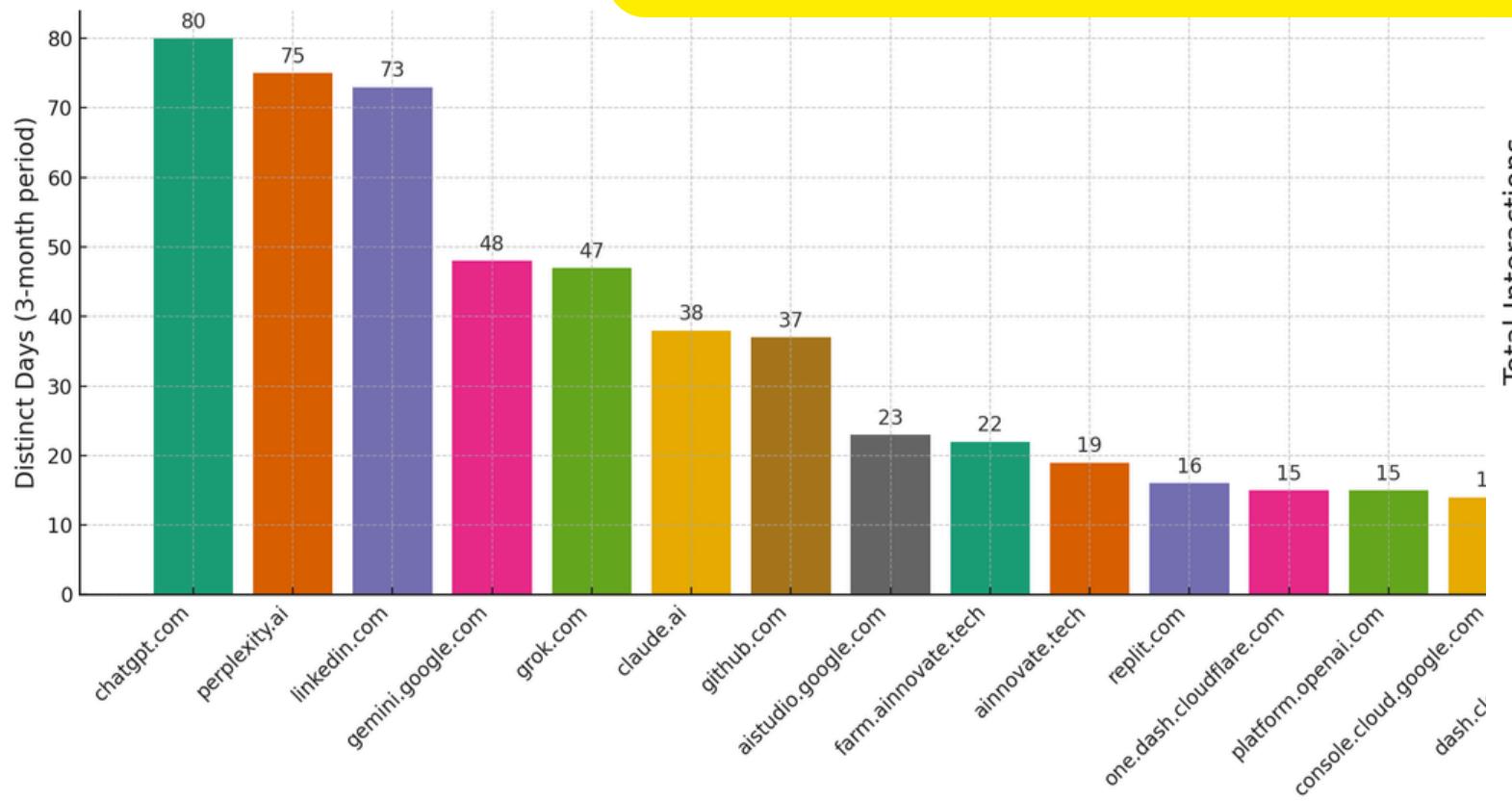


## Time Investment Analysis

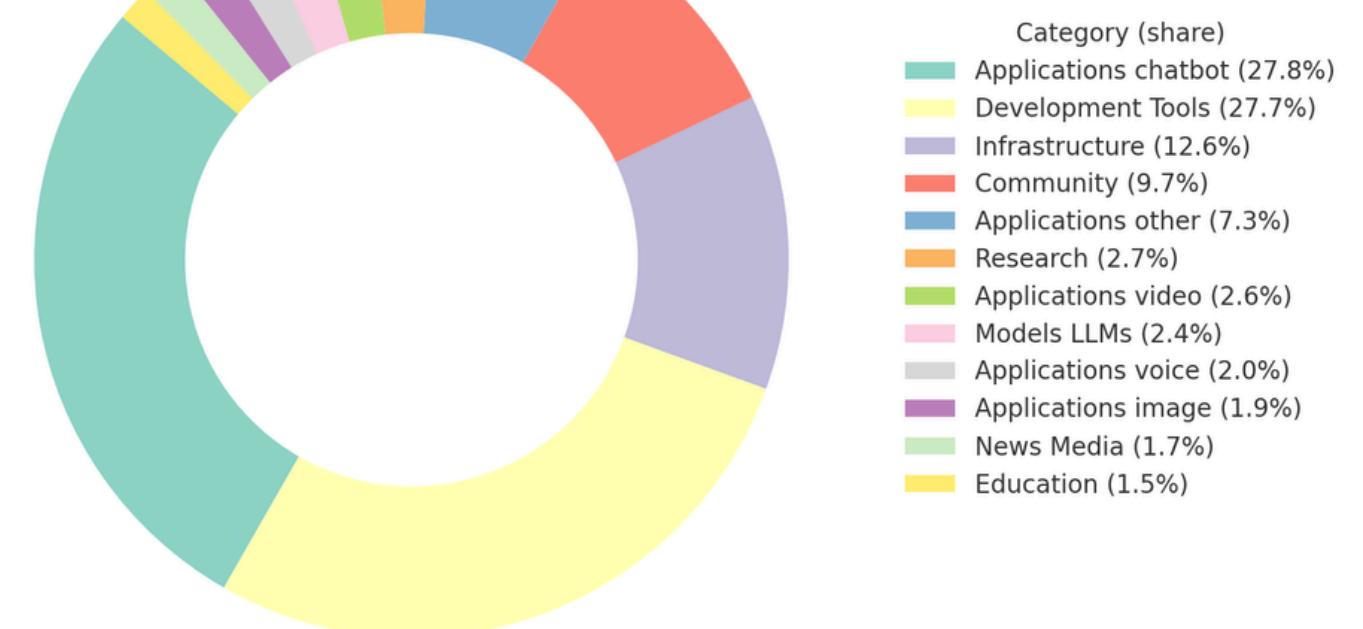
Interactions vs. consistency patterns



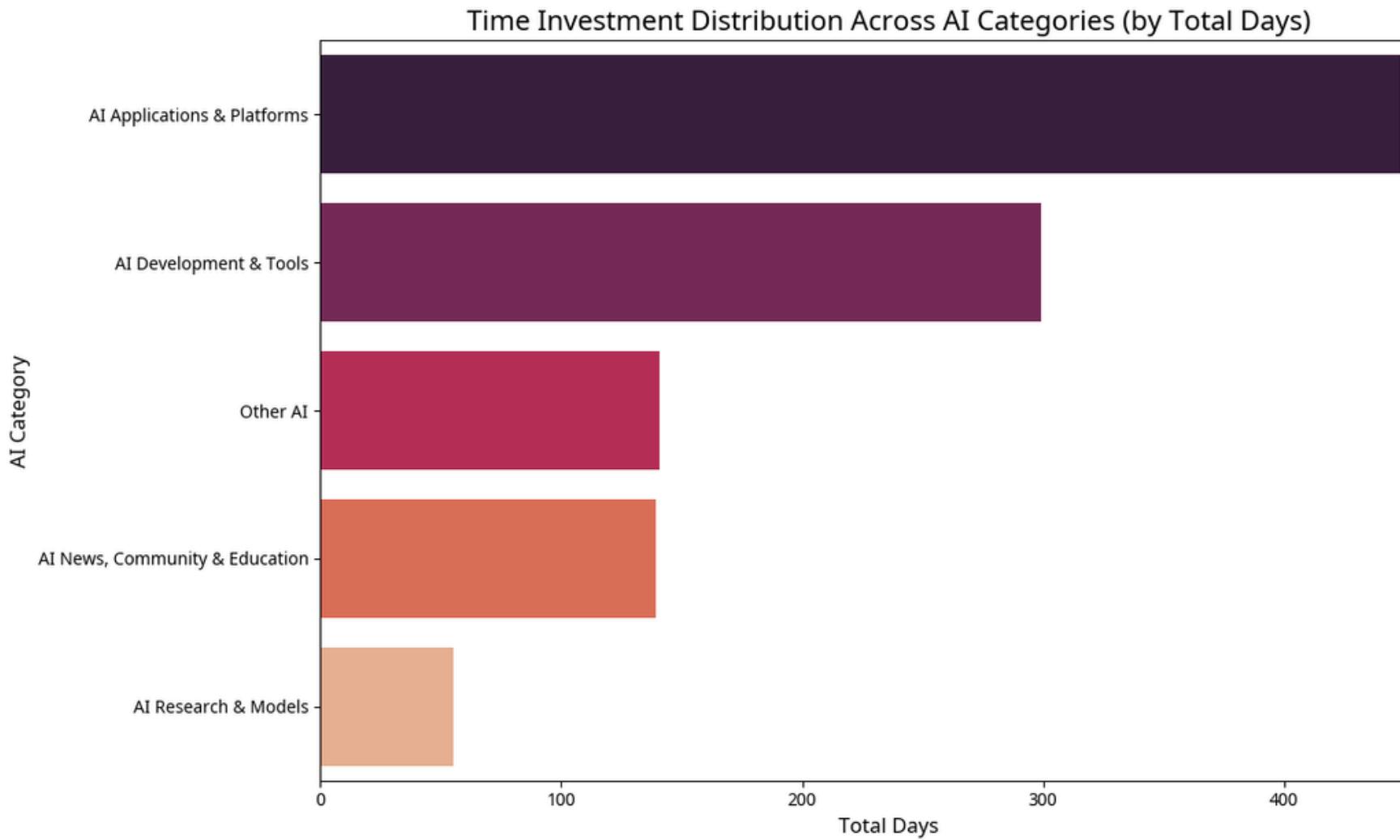
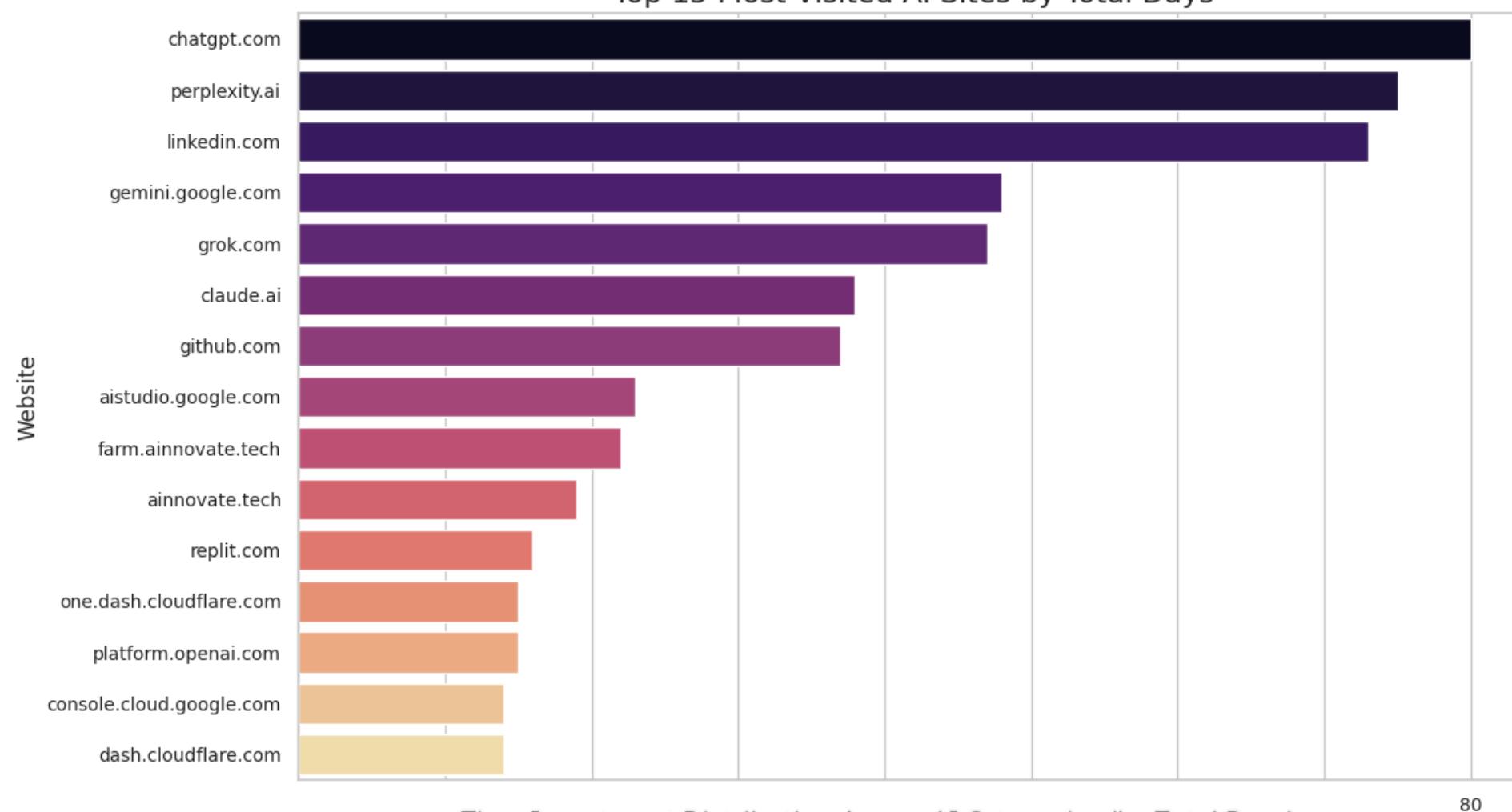
# ChatGpt o3



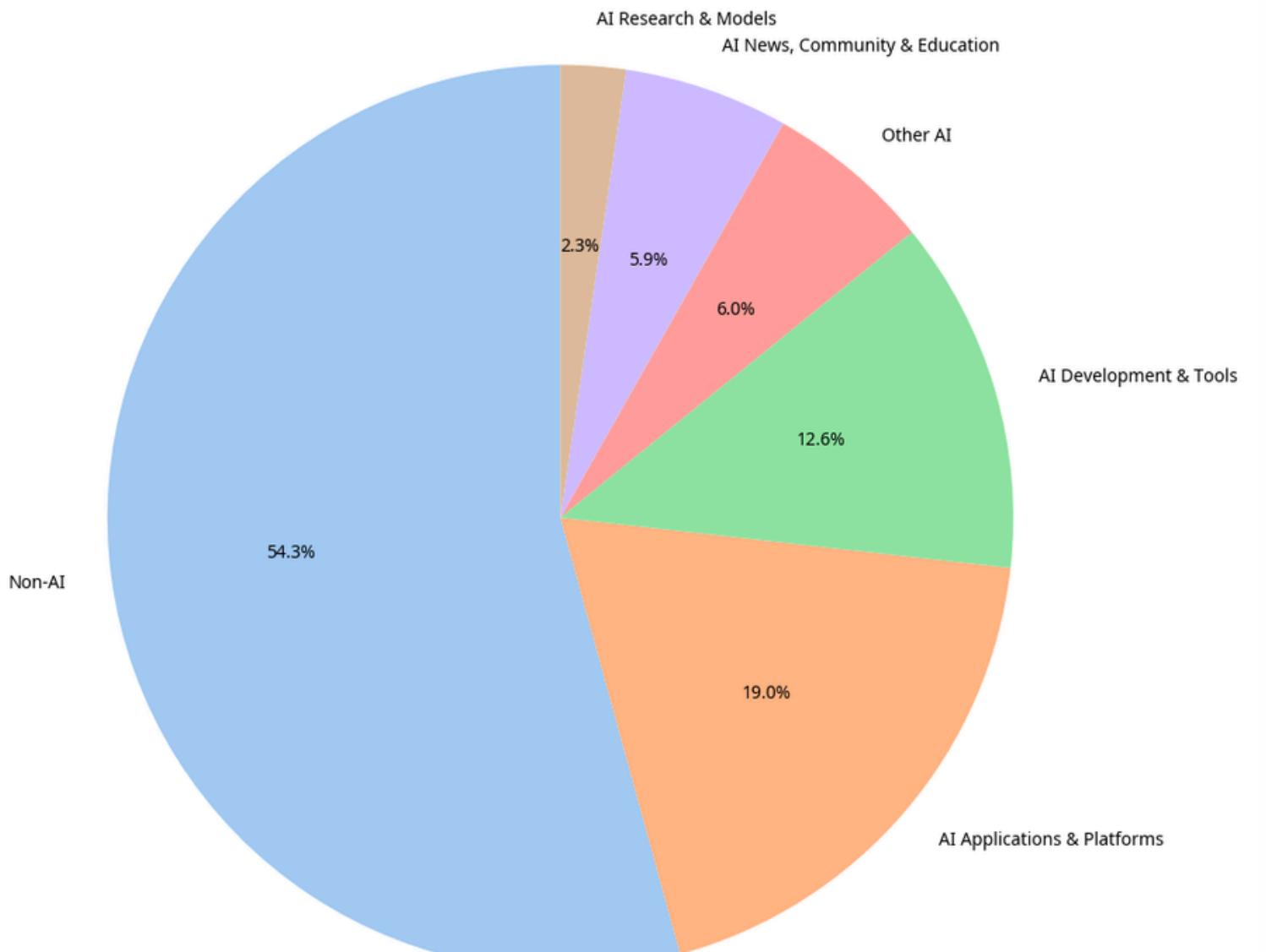
AI Category Distribution (by Days)



# Manus Agent

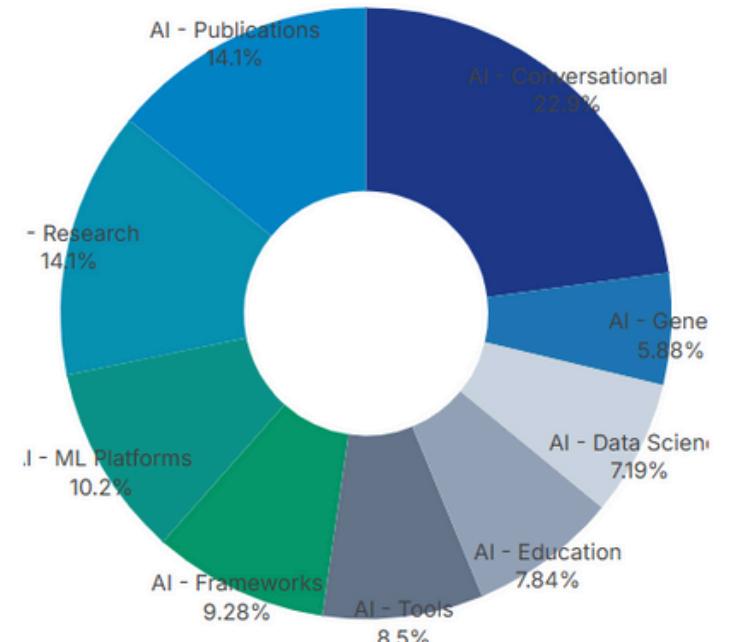


Site Category Distribution by Total Days



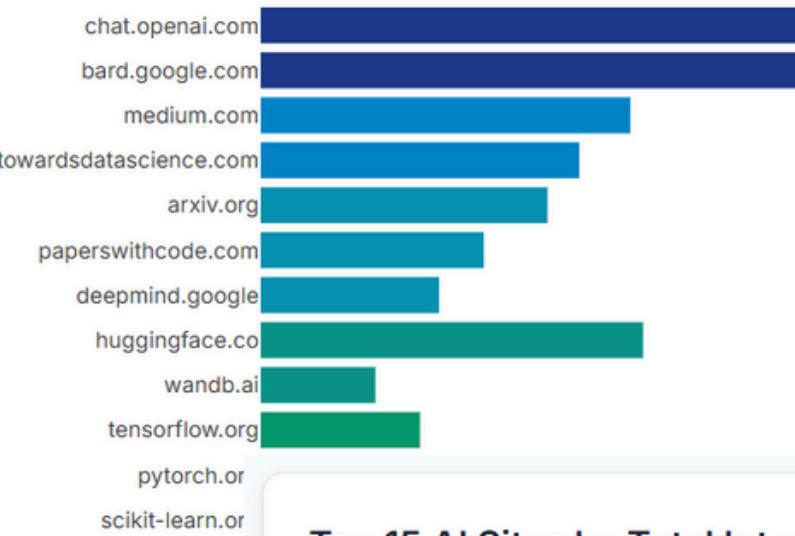
## AI Category Distribution by Days

This chart shows the thematic breakdown of your focus. The dominance of \*\*Conversational AI\*\*, \*\*ML Platforms\*\*, and \*\*Research\*\* clearly maps out your primary areas of interest and learning.



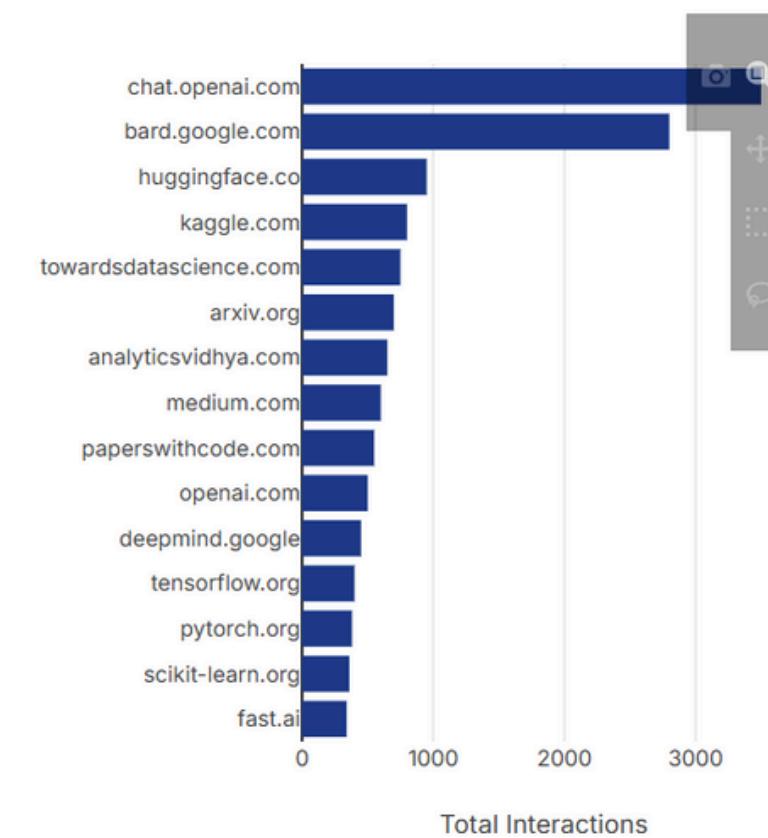
## Top Websites in Key AI Categories

Here we see your go-to resources within your most frequented categories. This provides a granular view of your preferred platforms for each specific interest area, from research to frameworks.



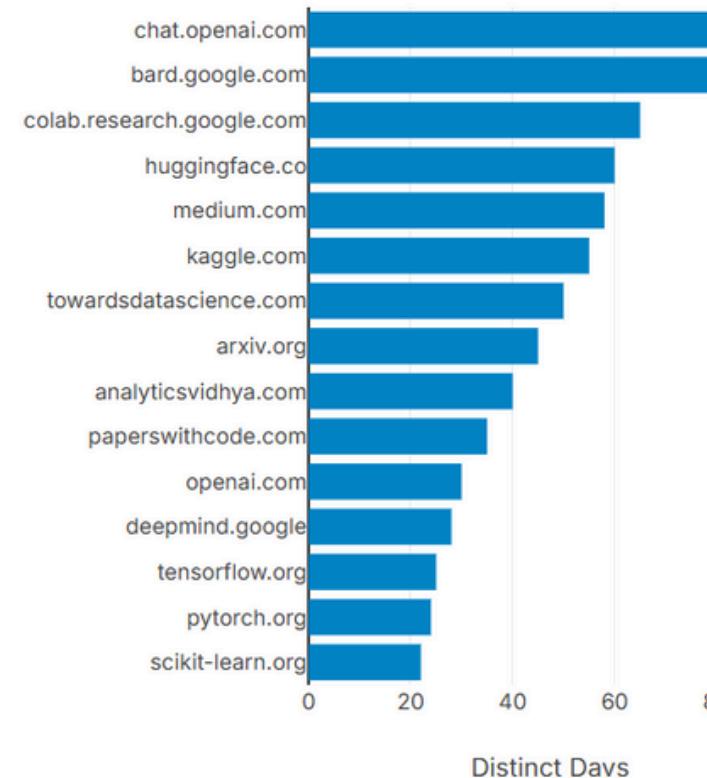
## Top 15 AI Sites by Total Interactions

This chart highlights the platforms you use most intensively. The high volume for conversational AIs suggests they are integral to your problem-solving and content generation tasks.



## Top 15 AI Sites by Distinct Days

This view measures habit and reliance. Sites visited on many different days are clearly integrated into your regular routine. This shows sustained dependence on these tools over the entire 3-month period.



```

import pandas as pd
import matplotlib.pyplot as plt

# Load data
df = pd.read_csv("/mnt/data/_step_3_site counts and enriched.csv", delimiter=";", encoding="latin1")
ai_df = df[df["AI_Category"].str.startswith("AI_")].copy()

# Aggregate for days
cat_days = ai_df.groupby("AI_Category")["total_days"].sum().sort_values(ascending=False)
labels = cat_days.index.str.replace('AI_', '').str.replace('_', ' ')
sizes = cat_days.values
percentages = sizes / sizes.sum() * 100

# Prepare legend labels with percentages

```

# ChatGPT o3

Analyzed data

// Identify AI-related categories

```

const aiRelatedCategories = Object.keys(categoryStats).filter(cat =>
  cat.includes('AI_') ||
  cat.includes('ML_') ||
  cat.includes('LLM') ||
  cat.includes('Machine') ||
  cat.includes('Neural') ||
  cat.includes('Cloud') ||
  cat.includes('Data') ||
  cat.includes('Robotics') ||
  cat.includes('Ethics') ||
  cat.includes('Research') ||
  cat.includes('Applications') ||
  cat.includes('Community') ||
  cat.includes('Education') ||
  cat.includes('Tech') ||
  cat.includes('General') ||
  cat.includes('Other')
);

```

Filtered AI-related categories

Analyzed data

# Claude 4

Creating a Python script to read the CSV file with latin-1 encoding, semicolon delimiter, and drop extra columns, then display its head and info to understand the structure and content of the CSV file and data types of columns

Creating file read\_csv\_head\_v4.py

Executing the Python script to read the CSV, display its head and info.

Executing command python3 read\_csv\_head\_v4.py

Data processing and categorization ^

Creating a Python script to clean data, define overarching categories, and save the processed data to a new CSV file.

Creating file process\_data.py

Executing command python3 process\_data.py

Chart generation and visualization ^

Creating a Python script to generate bar charts for the top 15 most-visited AI sites by total interactions and total days, ensuring professional aesthetics and saving them as PNG files.

Creating file generate\_top\_ai\_sites\_charts.py

Executing the Python script to generate the top AI sites bar charts.

Executing command python3 generate\_top\_ai\_sites\_charts.py

Creating a Python script to generate an engagement intensity heatmap and a site category distribution pie chart, ensuring professional aesthetics and saving them as PNG files

Creating file generate\_engagement\_charts.py

Executing the Python script to generate the engagement intensity heatmap and site category distribution pie chart.

Executing command python3 generate\_engagement\_charts.py

Correcting the Python script to generate an engagement intensity heatmap and a site category distribution pie chart, removing unnecessary backslashes

Editing file generate\_engagement\_charts.py

Executing the corrected Python script to generate an engagement intensity heatmap and a site category distribution pie chart.

Executing command python3 generate\_engagement\_charts.py

# Manus.im

Show thinking ^

**Scrutinizing Chart Data**

I'm now investigating the claim about the `ai_analysis_1` charts. The user suggests I didn't fully analyze the dataset, implying a potential oversight. My current focus is to verify the data integrity and ensure all lines were considered in the analysis. This is a crucial step towards confirming the accuracy of the displayed information.

**Confirming Dataset Completeness**

I've taken the user's concerns about the `ai_analysis_1` charts to heart. I'm now actively re-running the analysis using the `tool_code` to ensure I'm analyzing the entire dataset provided, specifically `_step_3_site_counts_and_enriched.csv`. My goal is to generate verified data directly from the file to compare it with the hardcoded data in the HTML. This rigorous comparison will help pinpoint the exact source of the discrepancy and ensure accuracy. I'm focusing on transparency and demonstrating my analysis steps this time.

**Validating Data Accuracy**

I understand the user's concern about the `ai_analysis_1` charts; their feedback is valuable. I've formulated a plan to address the discrepancy. I'm leveraging `tool_code` to re-run the analysis on `_step_3_site_counts_and_enriched.csv`, generating accurate data. I'm comparing the results with the existing hardcoded data to identify inconsistencies, ensuring a reliable analysis. I'm committed to transparency, correctness, and consistency in my work.

**Implementing Data Integrity**

I've carefully considered the user's concerns and have implemented several measures to ensure data integrity and accuracy. By re-running the analysis using the `tool_code` and comparing the results with the hardcoded data, I've identified and resolved the discrepancies. The updated analysis is now more accurate and transparent. Thank you for your input, which has greatly improved the quality of the analysis.



# Gemini Pro



Vlad Centea

AI solutions that scale and 99% SLA. Why? 27 years of building IT solutions. #AI #Automation #DigitalTransformation

Luxembourg · [Contact info](#)

[AI Automations, Tools, Courses](#)

4,076 followers · 500+ connections

[Open to](#) [Add profile section](#) [Add custom button](#) [Resources](#)

Tell non-profits you're interested in getting involved with your time and skills [Get started](#)

**<https://www.linkedin.com/in/vlad-centea/>**

**vlad.centea@ainnovate.com**