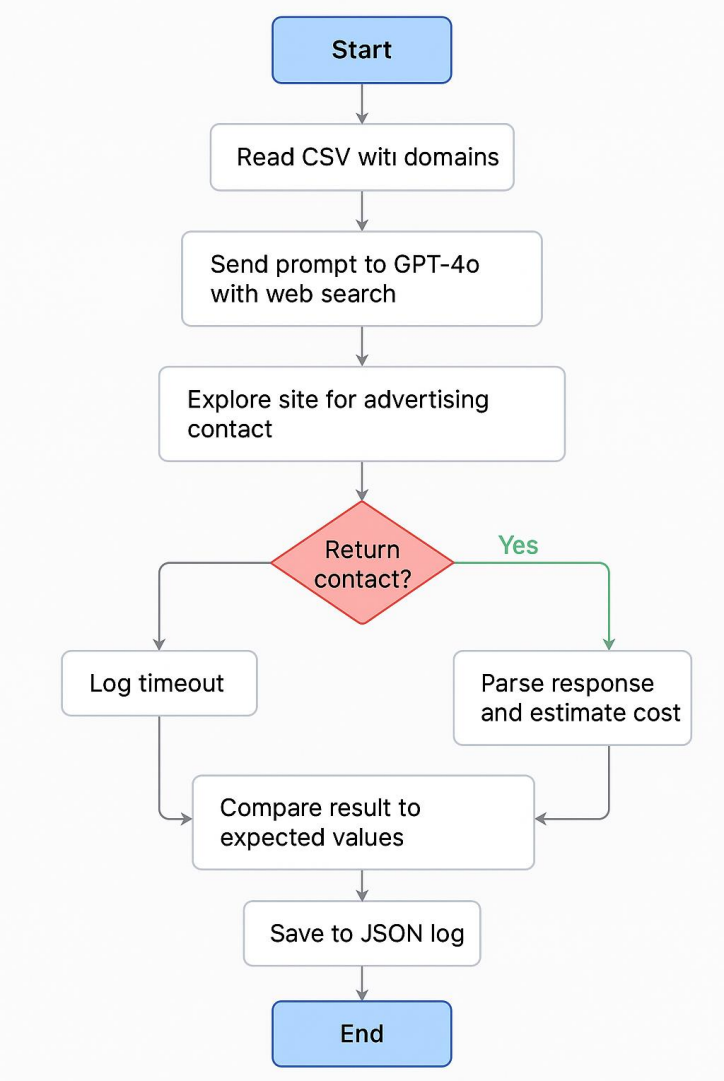
GPT-4o-mini Web Search

Total Domains Processed	**304**
Email Matching Success Rate	**17.43%**
Form Matching Success Rate	**0.33%**
Total Emails Found (any)	**231 (76%)**
Domains Timed Out	**0**
Total GPT Tokens Used	**212800**
Estimated GPT Cost	**$0.2128**
Total matched emails	**53 (17%)**
Total matched forms	**1 (0%)**
Total forms found	**86 (28%)**

Code Explanation:

This script automates the discovery of the best advertising-related contact method (email or form) for each domain listed in a CSV file using OpenAI's GPT-4o with web search. For every domain, it sends a detailed prompt instructing GPT to explore the site and return a single most relevant contact—prioritizing addresses like advertising@ or marketing forms. It handles JSON or regex-based parsing of the response, logs whether results match expected values and estimates token cost. Each result is saved as a JSON log, and a 2-minute timeout is enforced per domain to avoid stalls, with polite delays added to respect rate limits.



GPT-4o Web Search

Total Domains Processed	**304**
Email Matching Success Rate	**17.11%**
Form Matching Success Rate	**0.00%**
Total Emails Found (any)	**212 (70%)**
Domains Timed Out	**35**
Total GPT Tokens Used	**212800**
Estimated GPT Cost	**$1.0640**
Total matched emails	**52 (17%)**
Total matched forms	**0 (0%)**
Total forms found	**41 (13%)**

GPT-4.1 Web Search

Total Domains Processed	**246**
Email Matching Success Rate	**19.11%**
Form Matching Success Rate	**0.00%**
Total Emails Found (any)	**158 (64%)**
Domains Timed Out	**0**
Total GPT Tokens Used	**154004**
Estimated GPT Cost	**$0.7707**
Total matched emails	**47 (19%)**
Total matched forms	**0 (0%)**
Total forms found	**1 (0%)**

- GPT-4.1 and GPT-4o both have the same approach/logic as GPT-4o-Mini, with the only main difference being the change in API type.

Custom Automation (Scraping + GPT API)

Total Domains Processed	**294**
Manual Success Rate	**33.67%**
GPT Success Rate	**28.23%**
Manual Email Found (any)	**165 (56%)**
GPT Email Found (any)	**159 (54%)**
Domains Timed Out	**54**
Used Recovery Logic	**25**
Total GPT Tokens Used	**266189**
Estimated GPT Cost	**$1.2409**

Code Explanation:

High-Level Functionality

- Automates domain-by-domain analysis to find the best advertising, marketing, or press-related contact method (email or form).
- Uses Selenium to load pages, navigate links, and extract HTML content.
- Integrates OpenAI's GPT-4o-mini model to summarize text, assess relevance, and assist in intelligent form selection.
- Saves per-domain results in structured JSON log files including emails, form details, match status, and token usage.

Link Extraction & Navigation

- Extracts links from either just the <header>/<footer> or full <body> depending on settings or fallback attempts.
- Filters out:
- Empty or mailto: links.
- Links with misleading text (e.g., "about us" not starting with "about").
- Duplicate or previously seen URLs.
- Scores relevance of link text using NLP (spacy), checking for presence of intent keywords and absence of exclusion phrases.
- Navigates only to relevant subpages first; if none found, performs a fallback run over broader page content.

Page Text and Email Extraction

- Extracts visible text using BeautifulSoup and captures it per visited page.
- Performs both:
- Manual extraction of emails using regex, including deobfuscation patterns like [at], (dot).
- GPT-based summarization to extract emails and assess contextual relevance for marketing/contact.
- Handles token budgeting and text summarization when content exceeds allowed size (via a summarize_page_text step).
- Compares both GPT-found and regex-found emails with the expected value for evaluation.

Form Extraction and Evaluation

- Detects all <form> tags on a page and applies multiple pre-filters:
- Skips search forms with only one field or labeled as search.
- Requires presence of both a <textarea> (for message input) and at least one other field.
- Uses a variety of techniques to parse and label form fields, including:
- label[for=id] match
- Parent <label> tag
- Nearby div/span text
- Placeholder or aria-label
- Previous text node
- Infers field required status based on HTML attributes or visual cues like * in label.
- Removes hidden fields and handles textarea evaluation with GPT to pick the most meaningful one (if unrequired).

Form Relevance & Selection via GPT

- GPT is used to:
    - Evaluate if a form is explicitly for advertising/marketing and not just general contact.
    - Summarize the full form+page context into a short sentence.
    - Choose the most appropriate form out of multiple detected forms based on those summaries.
- Saves the final chosen form's HTML to a dedicated directory for inspection/debugging.