An AI-Powered Workflow To Solve Content Cannibalization via @sejournal, @Kevin_Indig

Your site likely suffers from at least some content cannibalization, and you might not even realize it.

Cannibalization hurts organic traffic and revenue: The impact can stretch from key pages not ranking to algorithm issues due to low domain quality.

However, cannibalization is tricky to detect, can change over time, and exists on a spectrum.

It's the "microplastics of SEO."

In this Memo, I'll show you:

1. How to identify and fix content cannibalization reliably.
2. How to automate content cannibalization detection.
3. An automated workflow you can try out right now: The Cannibalization Detector, my new keyword cannibalization tool.

I could have never done this without Nicole Guercia from AirOps. I've designed the concept and stress-tested the automated workflow, but Nicole built the whole thing.

How To Think About Content Cannibalization The Right Way

Before jumping into the workflow, we must clarify a few guiding principles about content cannibalization that are often misunderstood.

The biggest misconception about cannibalization is that it happens on the keyword level.

It's actually happening on the user intent level.

We all need to stop thinking about this concept as keyword cannibalization and instead as content cannibalization based on user intent.

With this in mind, cannibalization…

- **Is a moving target:** When Google updates its understanding of intent during a core update, suddenly two pages can compete with each other that previously didn't.
- **Exists on a spectrum:** A page can compete with another page or several pages, with an intent overlap from 10% to 100%. It's hard to say exactly how much overlap is fine without looking at outcomes and context.
- **Doesn't stop at rankings:** Looking for two pages that are getting a "substantial" amount of impressions or rankings for the same keyword(s) can help you spot cannibalization, but it is not a very accurate method. It's not enough proof.
- **Needs regular check-ups:** You need to check your site for cannibalization regularly and treat your content library as a "living" ecosystem.
- **Can be sneaky:** Many cases are not clear-cut. For example, international content cannibalization is not obvious. A /en directory to address all English-speaking countries can compete with a /en-us directory for the U.S. market.

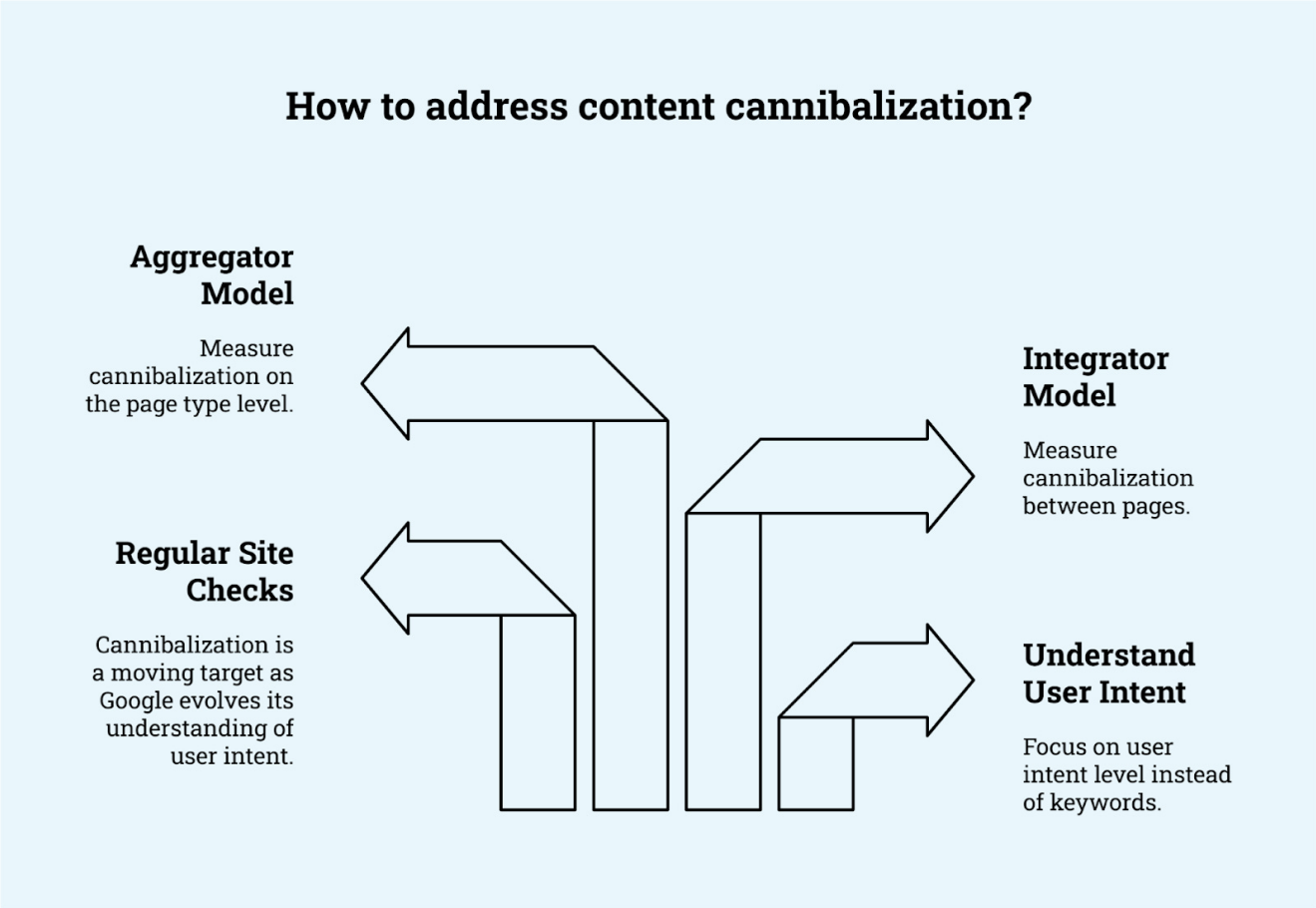


Image Credit: Kevin Indig

Different types of sites have fundamentally different weaknesses for cannibalization.

My model for site types is the integrator vs. aggregator model. Online retailers and other marketplaces face fundamentally different cases of cannibalization than SaaS or D2C companies.

Integrators cannibalize between pages. **Aggregators** cannibalize between page types.

- **With aggregators**, cannibalization often happens when two page types are too similar. For example, you can have two page types that could or could not compete with each other: “points of interest in {city}” and “things to do in {city}”.
- **With integrators**, cannibalization often happens when companies publish new content without maintenance and a plan for the existing content. A big part of the issue

is that it becomes harder to keep an overview of what you have and what keywords/intent it targets at a certain number of articles (I found the linchpin to be around 250 articles).

How To Spot Content Cannibalization



An example of content cannibalization (Image Credit: Kevin Indig)

Content cannibalization can have one or more of the following symptoms:

- "URL flickering": meaning at least two URLs alternate in ranking for one or more keywords.
- A page loses traffic and/or ranking positions after another one goes live.
- A new page hits a ranking plateau for its main keyword and cannot break into the top 3 positions.
- Google doesn't index a new page or pages within the same page type.
- Exact duplicate titles appear in Google's search index.
- Google reports "crawled, not indexed" or "discovered, not indexed" for URLs that don't have thin content or technical issues.

Since Google doesn't give us a clear signal for cannibalization, the best way to measure similarity between two or more pages is cosine similarity between their tokenized

embeddings (I know, it’s a mouthful).

But this is what it means: Basically, you compare how similar two pages are by turning their text into numbers and seeing how closely those numbers point in the same direction.

Think about it like a chocolate cookie recipe:

- **Tokenization** = Break down each recipe (e.g., page content) into ingredients: flour, sugar, chocolate chips, etc.
- **Embeddings** = Convert each ingredient into numbers, like how much of each ingredient is used and how important each one is to the recipe’s identity.
- **Cosine Similarity** = Compare the recipes mathematically. This gives you a number between 0 and 1. A score of 1 means the recipes are identical, while 0 means they’re completely different.

Follow this process to scan your site and find cannibalization candidates:

- **Crawl:** Scrape your site with a tool like Screaming Frog (optionally, exclude pages that have no SEO purpose) to extract the URL and meta title of each page
- **Tokenization:** Turn words in both the URL and title into pieces of words that are easier to work with. These are your tokens.
- **Embeddings:** Turn the tokens into numbers to do “word math.”
- **Similarity:** Calculate the cosine similarity between all URLs and meta titles

Ideally, this gives you a shortlist of URLs and titles that are too similar.

In the next step, you can apply the following process to make sure they truly cannibalize each other:

- **Extract content:** Clearly isolate the main content (exclude navigation, footer, ads, etc.). Maybe clean up certain elements, like stop words.
- **Chunking or tokenization:** Either split content into meaningful chunks (sentences or paragraphs) or tokenize directly. I prefer the latter.
- **Embeddings:** Embed the tokens.
- **Entities:** Extract named entities from the tokens and weigh them higher in embeddings. In essence, you check which embeddings are “known things” and give them more power in your analysis.
- **Aggregation of embeddings:** Aggregate token/chunk embeddings with a weighted averaging (eg, TF-IDF) or attention-weighted pooling.
- **Cosine similarity:** Calculate cosine similarity between resulting embeddings.

You can use my app script if you’d like to try it out in Google Sheets (but I have a better alternative for you in a moment).

About cosine similarity: It’s not perfect, but good enough.

Yes, you can fine-tune embedding models for specific topics.

And yes, you can use advanced embedding models like sentence transformers on top, but this simplified process is usually sufficient. No need to make an astrophysics project out of it.

How To Fix Cannibalization

Once you’ve identified cannibalization, you should take action.

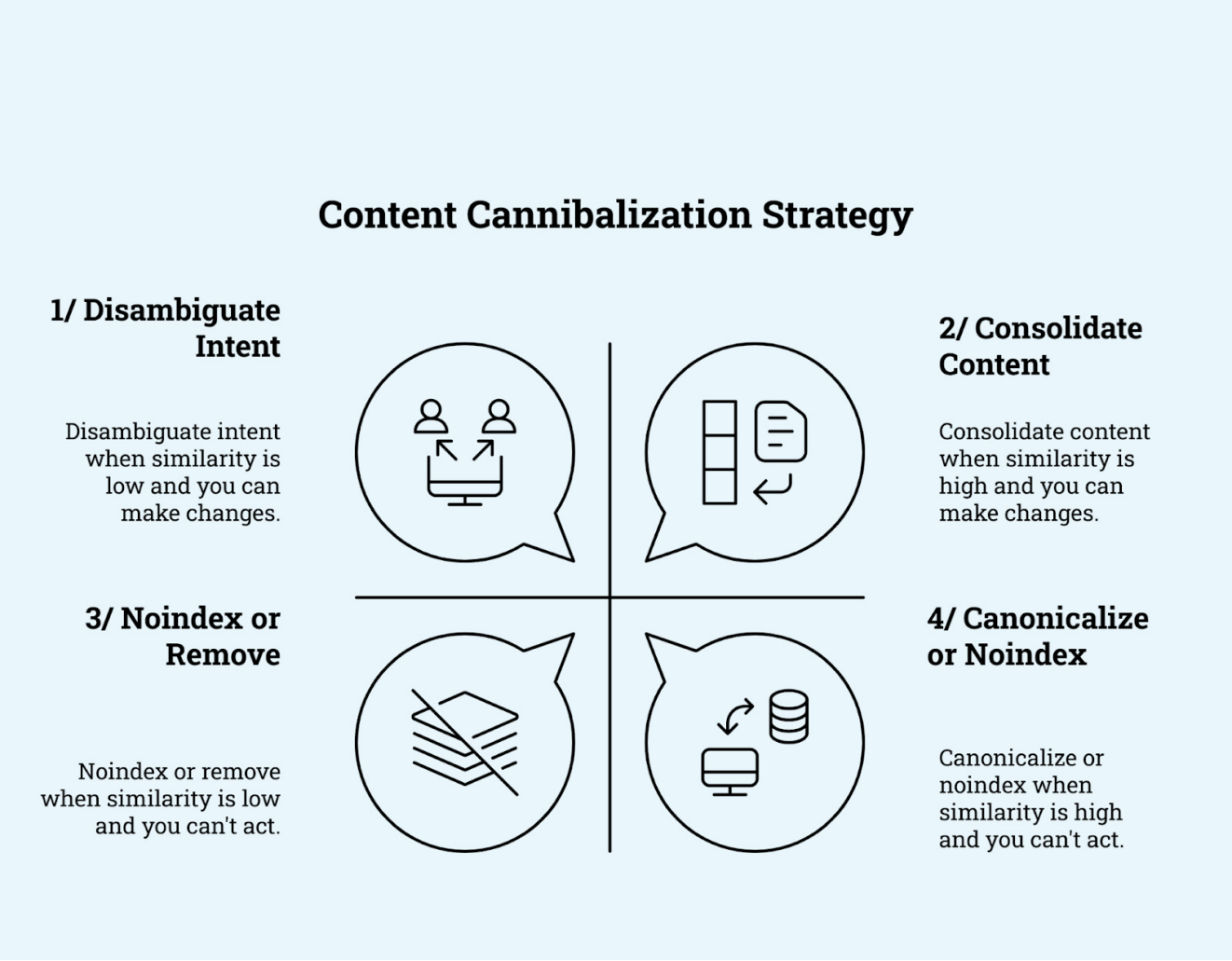
But don’t forget to adjust your long-term approach to content creation and governance. If you don’t, all this work to find and fix cannibalization is going to be a waste.

Solving Cannibalization In The Short Term

The short-term action you should take depends on the degree of cannibalization and how quickly you can act.

"Degree" means how similar the content across two or more pages is, expressed in cosine or content similarity.

Though not an exact science, in my experience, a cosine similarity higher than 0.7 is classified as "high", while it's "low" below a value of 0.5.



4 ways to fix cannibalization (Image Credit: Kevin Indig)

What to do if the pages have a high degree of similarity:

- **Canonicalize or noindex** the page when cannibalization happens due to technical issues like parameter URLs, or if the cannibalizing page is irrelevant for SEO, like paid landing pages. In this case, canonicalize the

parameter URL to the non-parameter URL (or noindex the paid landing page).
- **Consolidate** with another page when it's not a technical issue. Consolidation means combining the content and redirecting the URLs. I suggest taking the older page and/or the worse-performing page and redirecting to a new, better page. Then, transfer any useful content to the new variant.

What to do if the pages have a low degree of similarity:

- **Noindex or remove** (status code: 410) when you don't have the capacity or ability to make content changes.
- **Disambiguate** the intent focus of the content if you have the capacity, and if the overlap is not too strong. In essence, you want to differentiate the parts of the pages that are too similar.

Solving Cannibalization In The Long Term

It's critical to take long-term action to adjust your strategy or production process because content cannibalization is a symptom of a bigger issue, not a root cause.

(Unless we're talking about Google changing its understanding of intent during a core algorithm update, and that has nothing to do with you or your team.)

The most critical long-term changes you need to make are:

1. **Create a content roadmap:** SEO Integrators should maintain a living spreadsheet or database with all SEO-relevant URLs and their main target keywords and intent to tighten editorial oversight. Whoever is in charge of the content roadmap needs to ensure there is no overlap between articles and other page types. Writers need to have a clear target intent for new and existing content.
2. **Develop clear site architecture:** The pendant of a content map for SEO Aggregators is a site architecture

map, which is simply an overview of different page types and the intent they target. It's critical to underline the intent as you define it with example keywords that you verify on a regular basis ("Are we still ranking well for those keywords?") to match it against Google's understanding and competitors.

The last question is: *"How do I know when content cannibalization is fixed?"*

The answer is when the symptoms mentioned in the previous chapter go away:

- Indexing issues resolve.
- URL flickering goes away.
- No duplicate titles appear in Google's search index.
- "Crawled, not indexed" or "discovered, not indexed" issues decrease.
- Rankings stabilize and break through a plateau (if the page has no other apparent issues).

And, after working with my clients under this manual framework for years, I decided it's time to automate it.

Introducing: A Fully Automated Cannibalization Detector

Together with Nicole, I used AirOps to build a fully automated AI workflow that goes through 37 steps to detect cannibalization within minutes.

It performs a thorough analysis of content cannibalization by examining keyword rankings, content similarity, and historical data.

Below, I'll break down the most important steps that it automates on your behalf:

1. Initial URL Processing

The workflow extracts and normalizes the domain and brand name from the input URL.

This foundational step establishes the target website's identity and creates the baseline for all subsequent analysis.

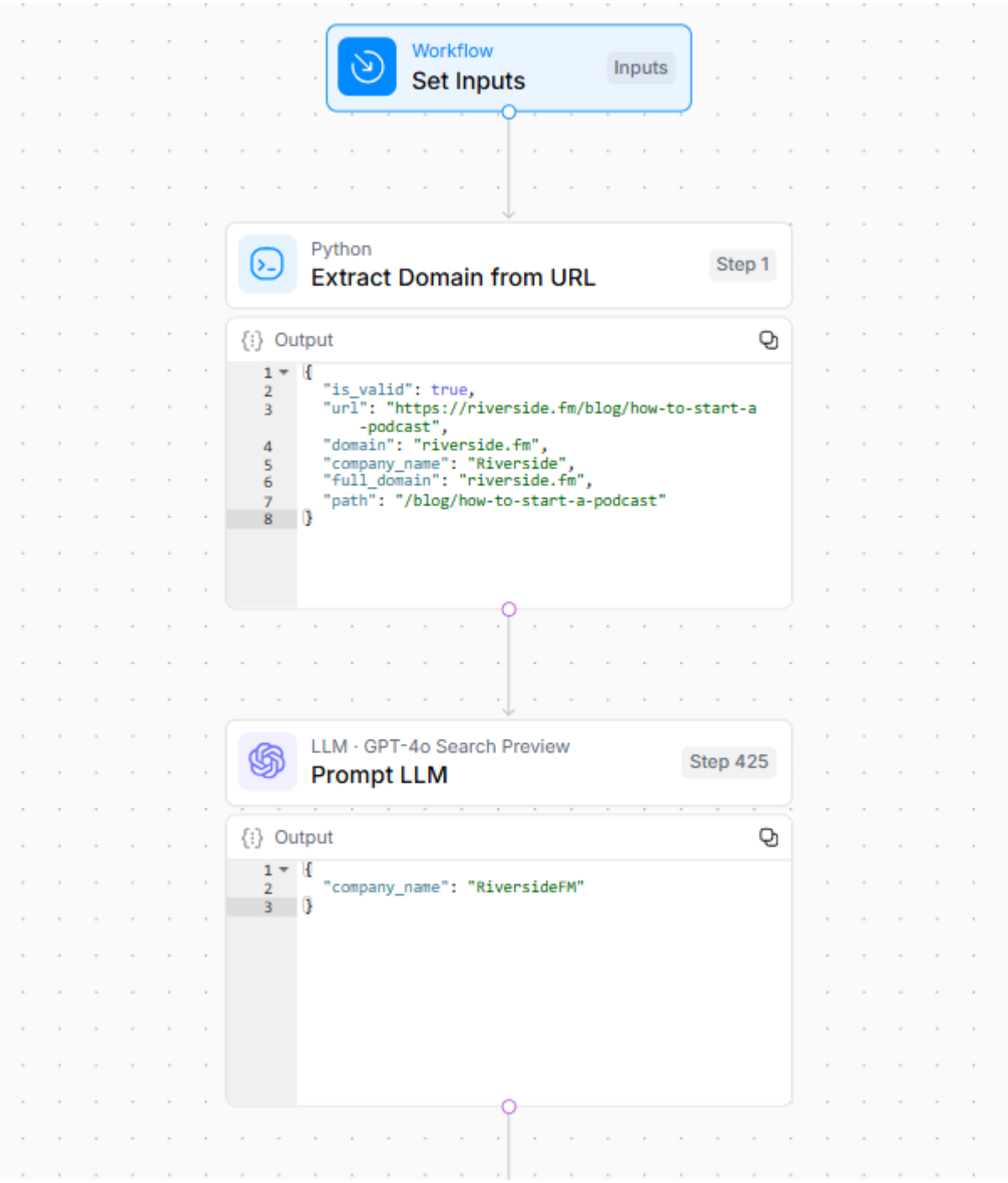


Image Credit: Kevin Indig

2. Target Content Analysis

To ensure that the system has quality source material to analyze and compare against competitors, Step 2 involves:

- Scraping the page.
- Validating and analyzing the HTML structure for main content extraction.
- Cleaning the article content and generating target embeddings.

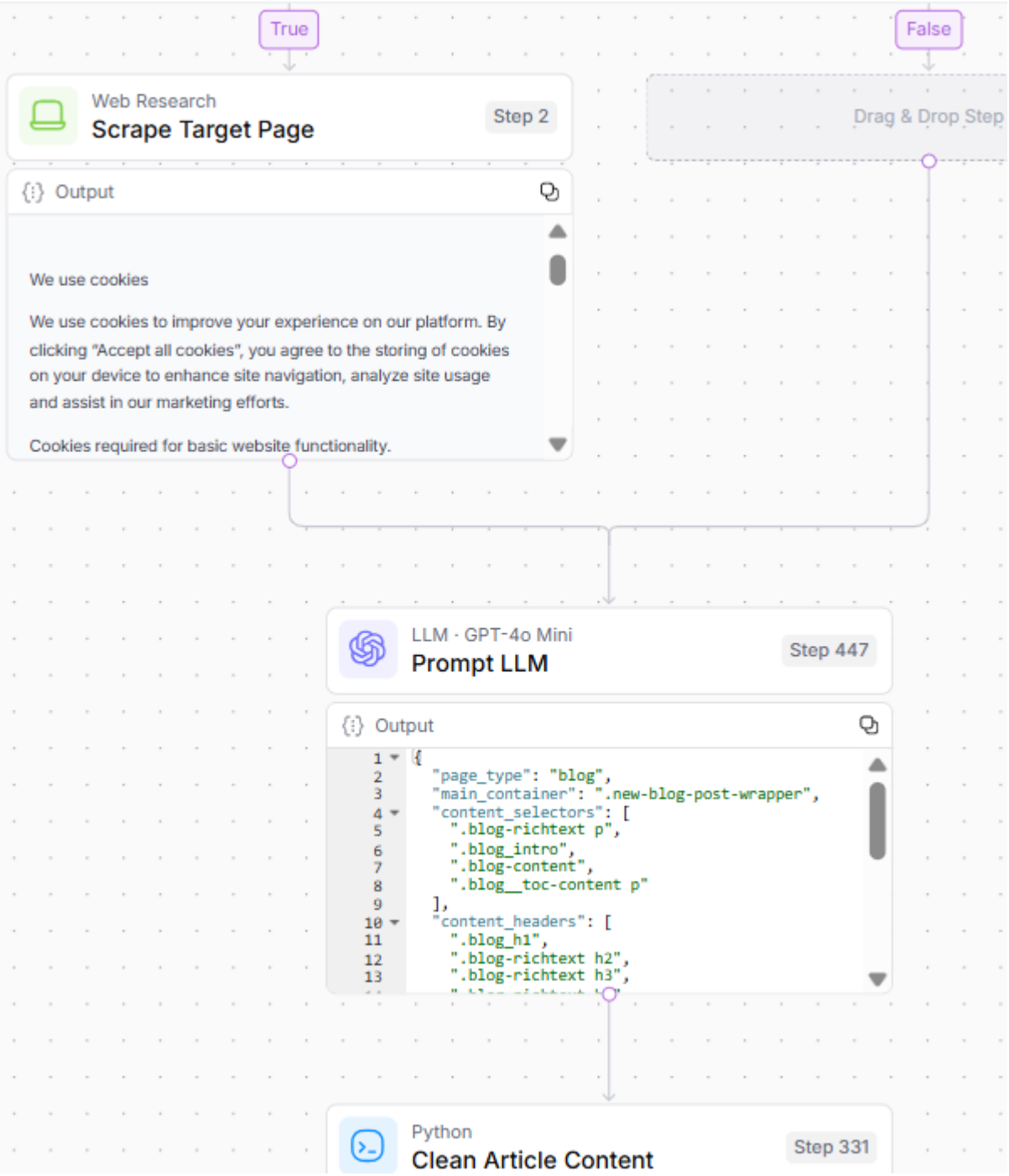


Image Credit: Kevin Indig

3. Keyword Analysis

Step 3 reveals the target URL's search visibility and potential vulnerabilities by:

- Analyzing ranking keywords through Semrush data.
- Filtering branded versus non-branded terms.
- Identifying SERP overlap with competing URLs.
- Conducting historical ranking analysis.
- Determining page value based on multiple metrics.
- Analyzing position differential changes over time.

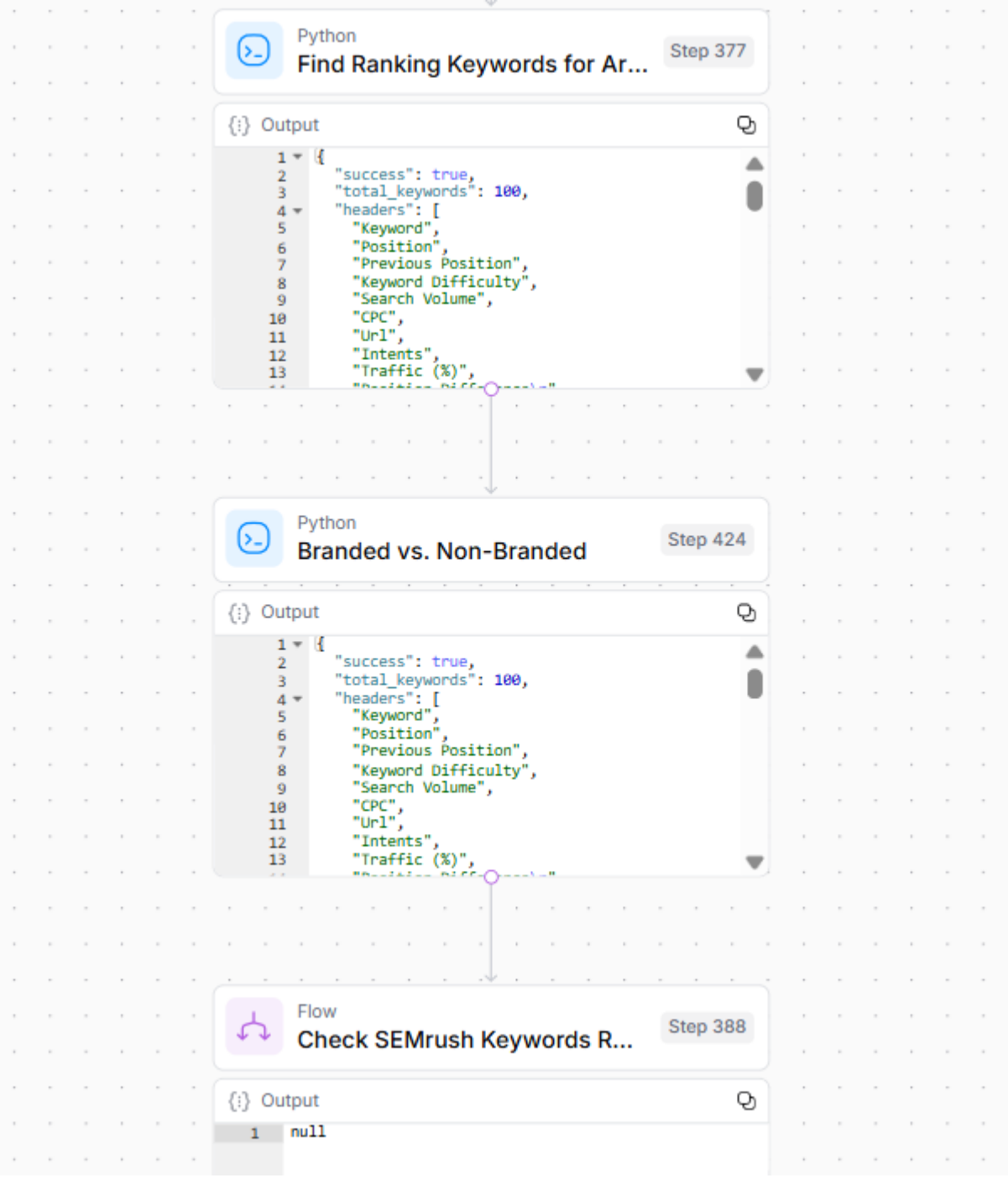


Image Credit: Kevin Indig

4. Competing Content Analysis (Iteration Over Competing URLs)

Step 4 gathers additional context for cannibalization by

iteratively processing each competing URL in the search results through the previous steps.

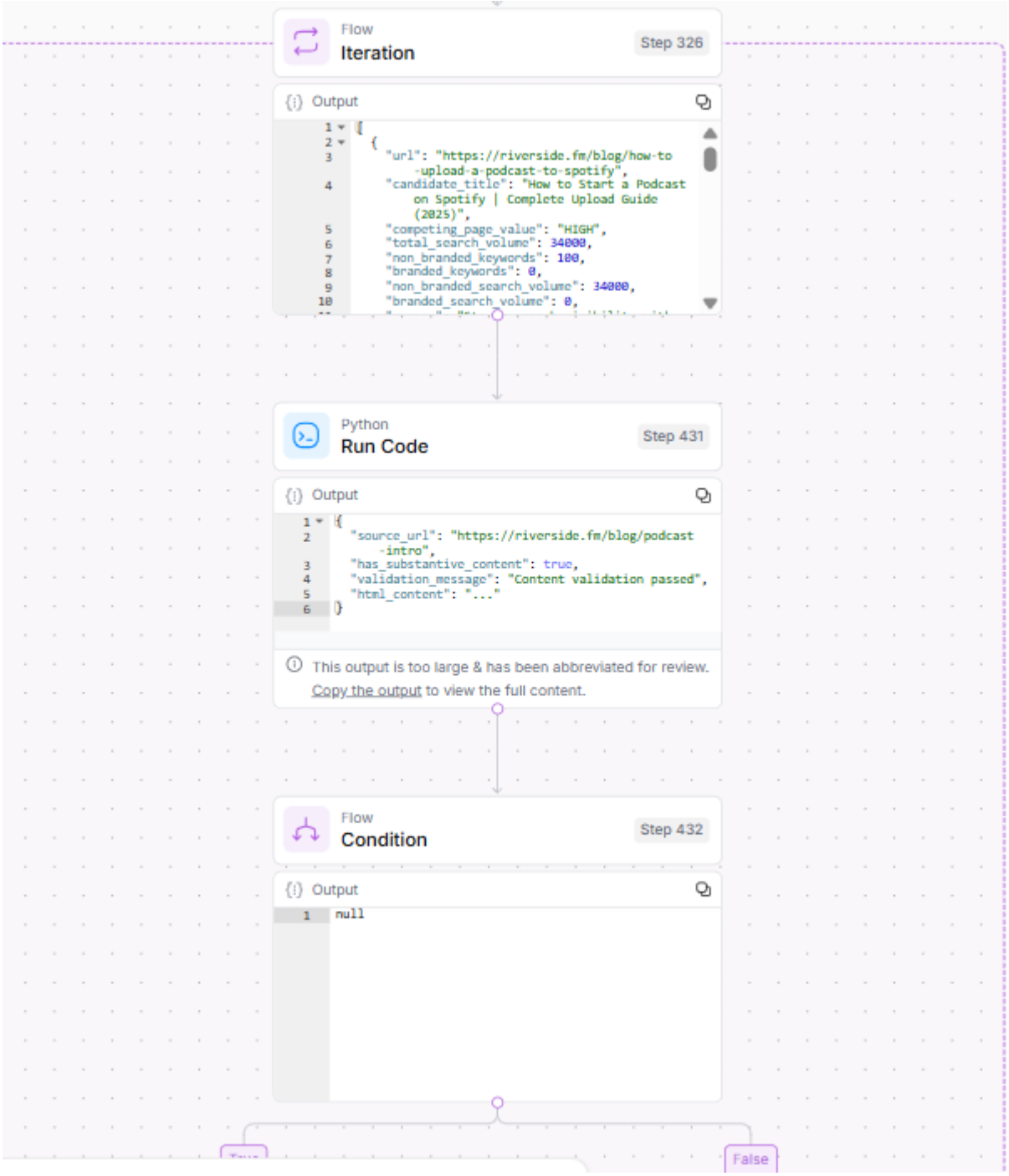


Image Credit: Kevin Indig

5. Final Report Generation

In the final step, the workflow cleans up the data and generates an actionable report.

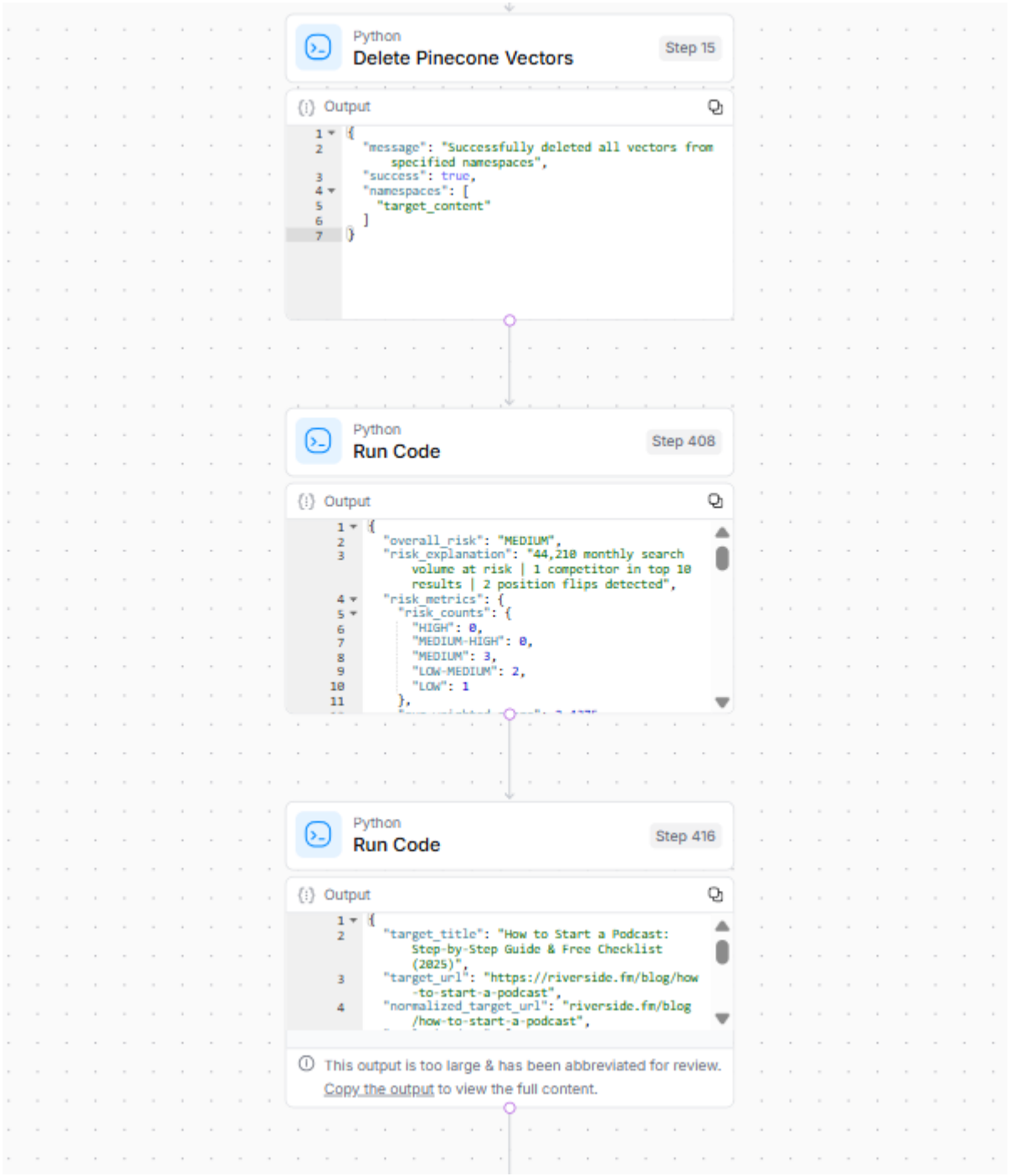
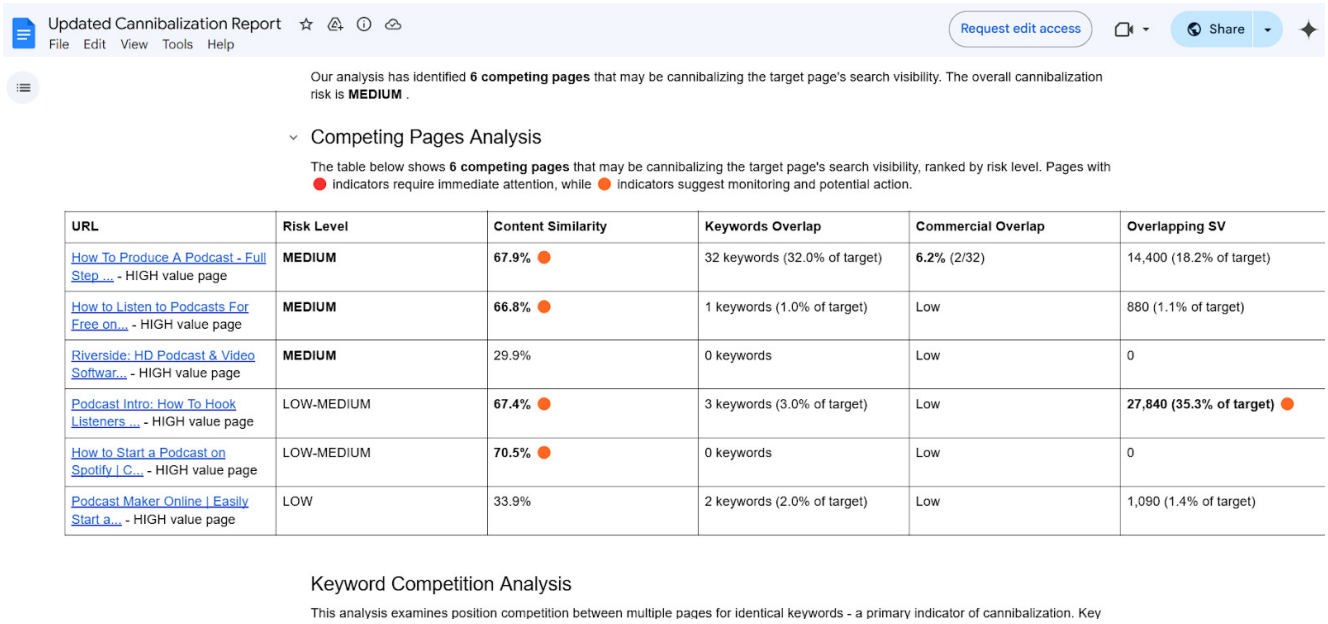


Image Credit: Kevin Indig

Try The Automated Content Cannibalization Detector



Updated Cannibalization Report

File Edit View Tools Help

Request edit access | Share

Our analysis has identified **6 competing pages** that may be cannibalizing the target page's search visibility. The overall cannibalization risk is **MEDIUM** .

Competing Pages Analysis

The table below shows **6 competing pages** that may be cannibalizing the target page's search visibility, ranked by risk level. Pages with 🔴 indicators require immediate attention, while 🟠 indicators suggest monitoring and potential action.

URL	Risk Level	Content Similarity	Keywords Overlap	Commercial Overlap	Overlapping SV
How To Produce A Podcast - Full Step ... - HIGH value page	**MEDIUM**	**67.9%** 🟠	32 keywords (32.0% of target)	**6.2%** (2/32)	14,400 (18.2% of target)
How to Listen to Podcasts For Free on... - HIGH value page	**MEDIUM**	**66.8%** 🟠	1 keywords (1.0% of target)	Low	880 (1.1% of target)
Riverside: HD Podcast & Video Softwar... - HIGH value page	**MEDIUM**	29.9%	0 keywords	Low	0
Podcast Intro: How To Hook Listeners ... - HIGH value page	LOW-MEDIUM	**67.4%** 🟠	3 keywords (3.0% of target)	Low	**27,840 (35.3% of target)** 🟠
How to Start a Podcast on Spotify	C... - HIGH value page	LOW-MEDIUM	**70.5%** 🟠	0 keywords	Low	0
Podcast Maker Online	Easily Start a... - HIGH value page	LOW	33.9%	2 keywords (2.0% of target)	Low	1,090 (1.4% of target)

Keyword Competition Analysis

This analysis examines position competition between multiple pages for identical keywords - a primary indicator of cannibalization. Key

Image Credit: Kevin Indig

Try the Cannibalization Detector and check out an example report.

A few things to note:

1. **This is an early version.** We're planning to optimize and improve it over time.
2. **The workflow can time out** due to a high number of requests. We intentionally limit usage so as not to get overwhelmed by API calls (they cost money). We'll monitor usage and might temporarily raise the limit, which means if your first attempt isn't successful, try again in a few minutes. It might just be a temporary spike in usage.
3. **I'm an advisor to AirOps** but was neither paid nor incentivized in any other way to build this workflow.

Please leave your feedback in the comments.

We'd love to hear how we can take the Cannibalization Detector to the next level!

Boost your skills with Growth Memo's weekly expert insights. Subscribe for free!

Featured Image: Paulo Bobita/Search Engine Journal

https://www.searchenginejournal.com/ai-workflow-content-cannibalization/543352/