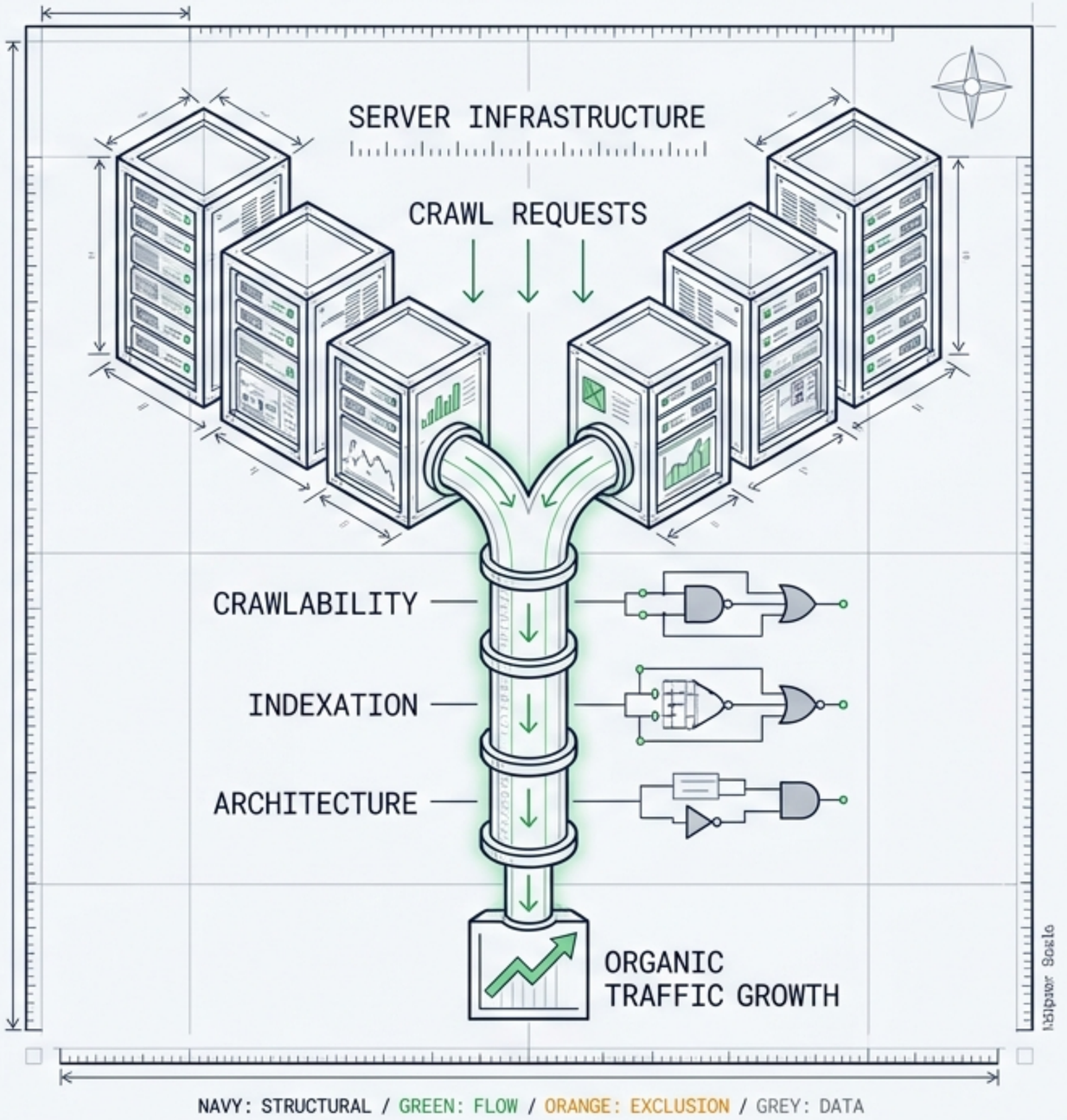
Technical SEO Fundamentals • Course 01

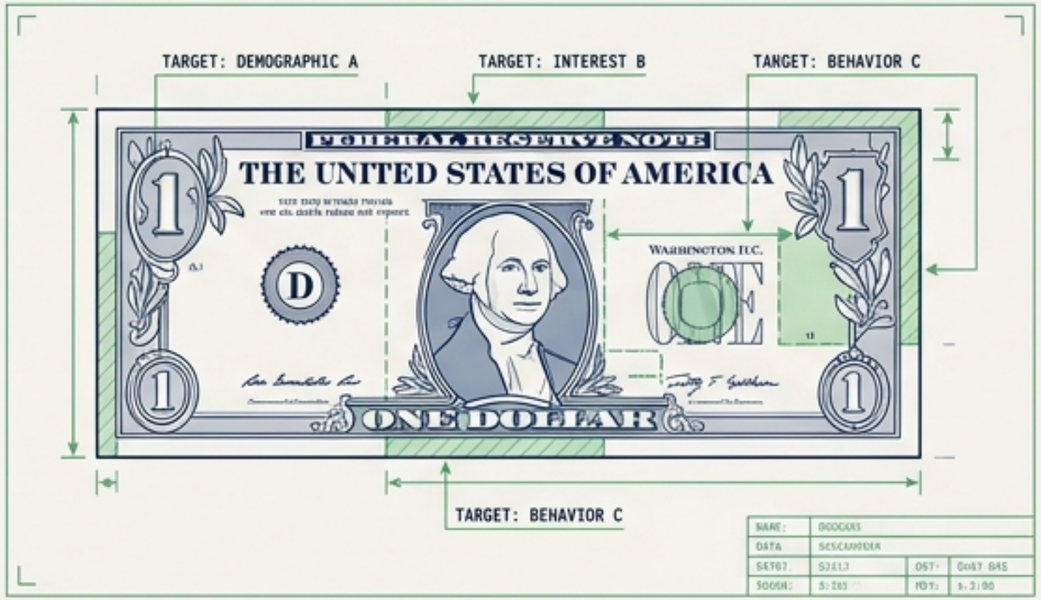
Organic Traffic Control: Directing the Bot Funnel

Crawlability, Indexation & Site Architecture Fundamentals for Large-Scale Sites



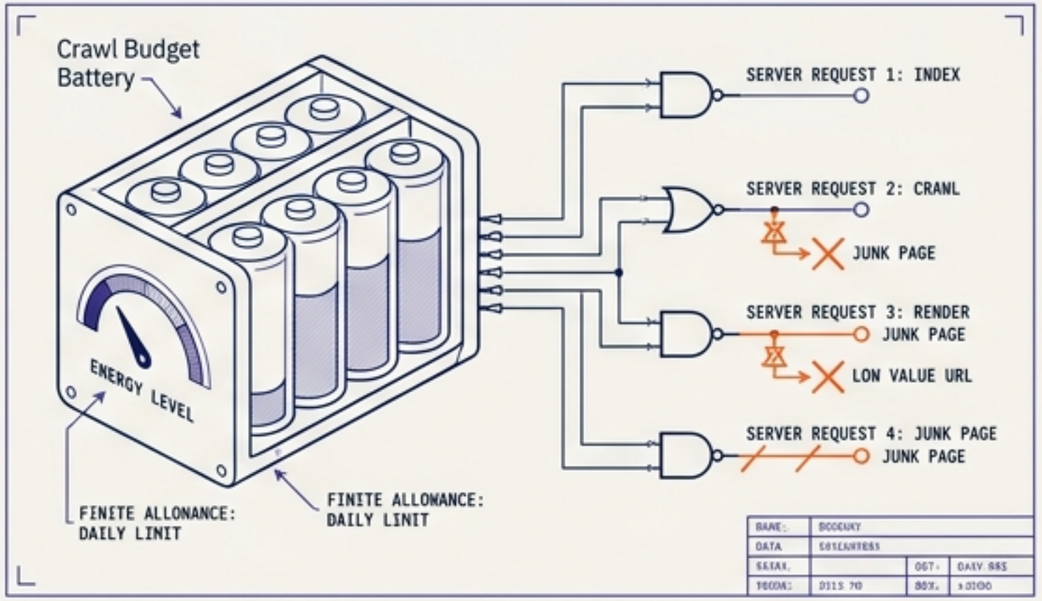
You already know how to manage a budget.

Paid Acquisition



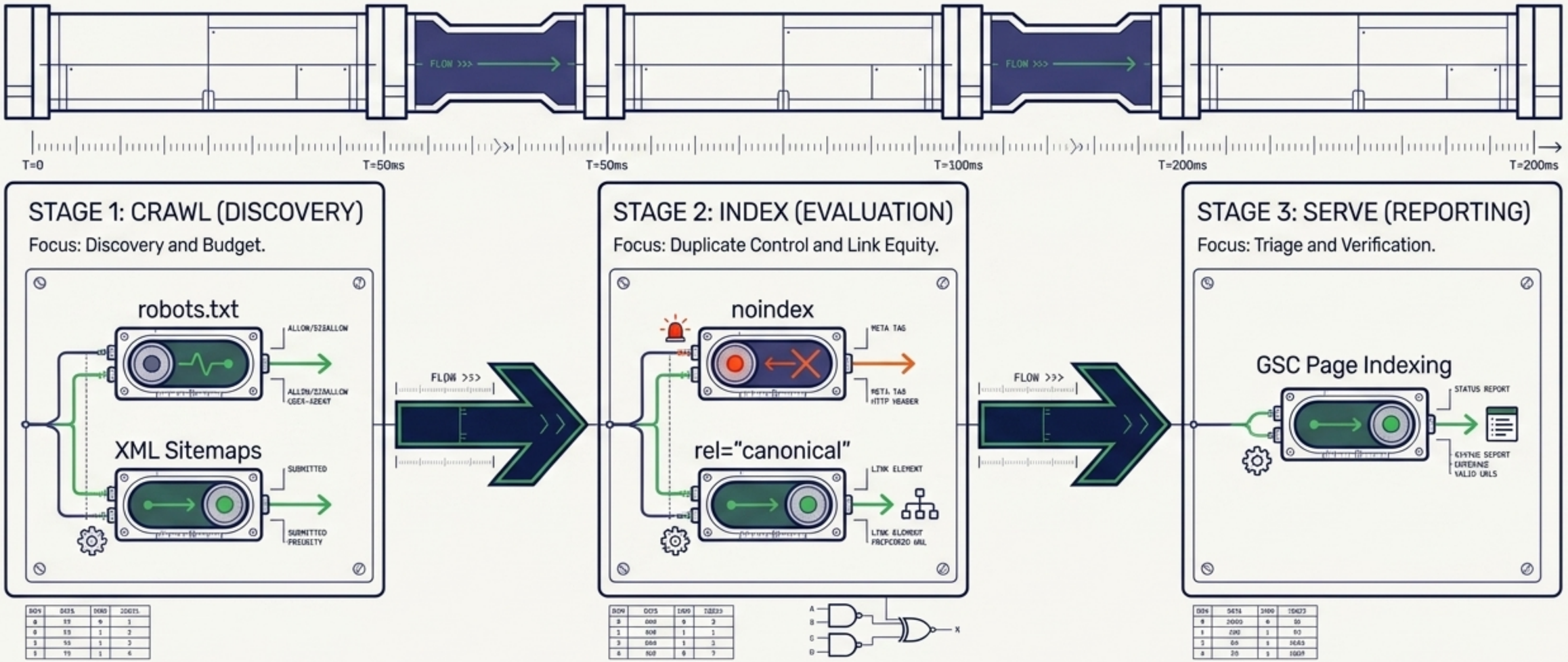
“Ad Spend” limits. You **exclude** bad demographics to stop **wasting dollars** on junk clicks.

Organic Acquisition



“Crawl Budget” limits. Googlebot has a **finite daily allowance** for your site. You exclude bad URLs to stop wasting crawl capacity on junk pages.

The Search Engine Pipeline



The Crawl Budget Equation

[Crawl Capacity] × [Crawl Demand] = Crawl Budget

Component 1: **Capacity**

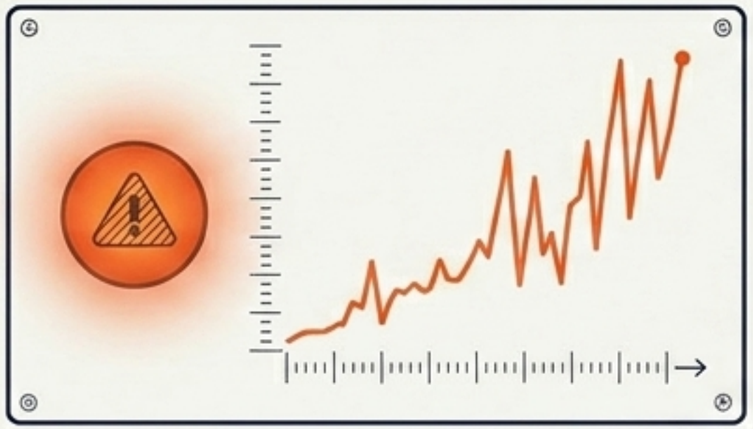
Determined by server responsiveness and Google's internal resources.

(You can reduce this in GSC, but never force an increase)

Component 2: **Demand**

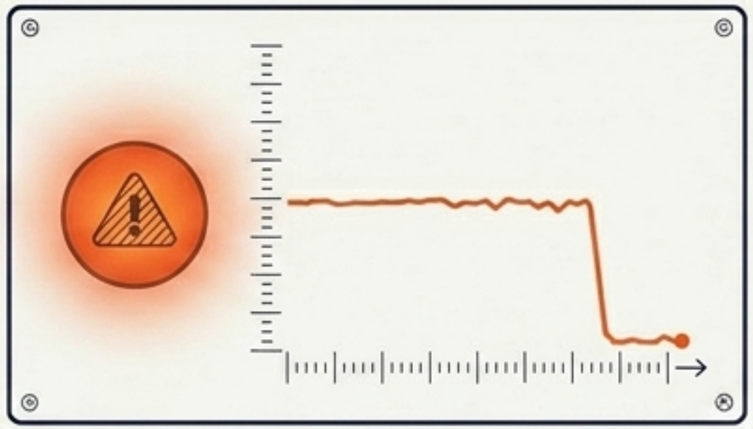
Driven by URL popularity, freshness, and perceived value.

Diagnosing Crawl Budget Exhaustion



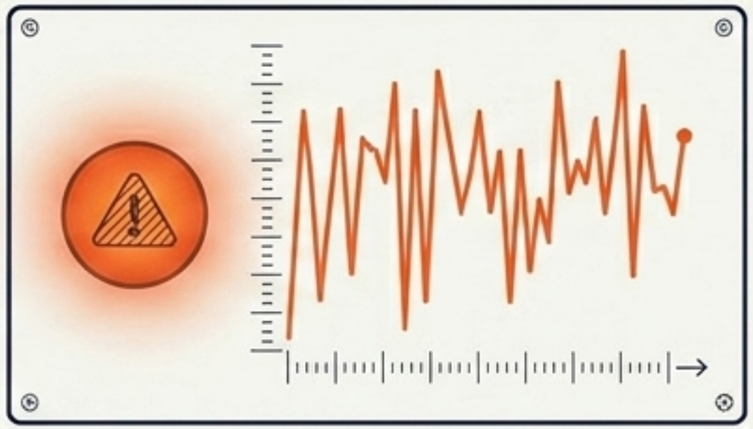
Warning 1: The GSC Backlog

Growing “Discovered — currently not indexed” count without corresponding Error reasons.



Warning 2: The Index Lag

Newly published pages (e.g., new destination guides) take `7+ days` to appear in URL Inspection.



Warning 3: The Junk Trap

Server logs show Googlebot spending high percentage of time crawling `/cart/`, `/checkout/`, or **session-parameterized** URLs.

The 2MB Fetch Limit

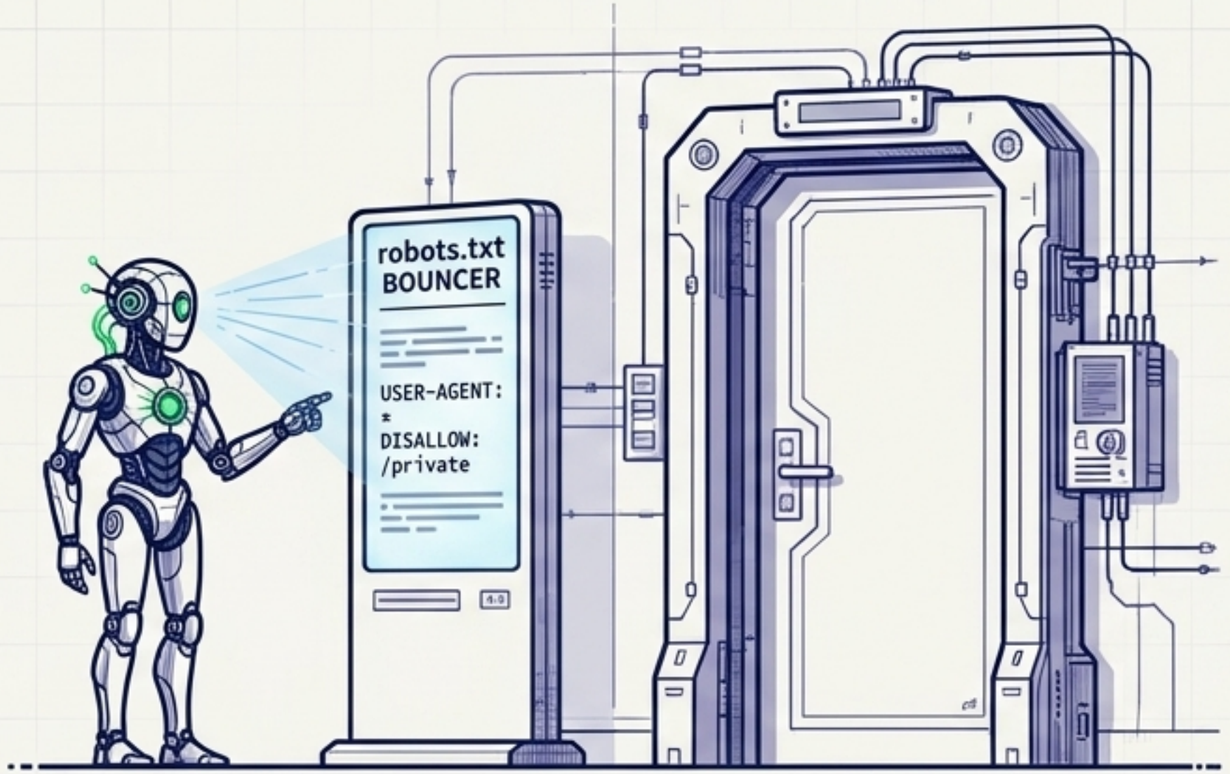
Data Point: Googlebot fetches a maximum of 2 MB of HTML per URL (excluding PDFs).

- **The Impact:** Content beyond this threshold is silently dropped.

OTA Use Case Alert

Long Single-Page Applications (SPAs) loading massive inline JSON objects for flight or hotel inventories risk having critical footer links or content severed from the index.

robots.txt is a Hint, Not a Lock



Core Principle

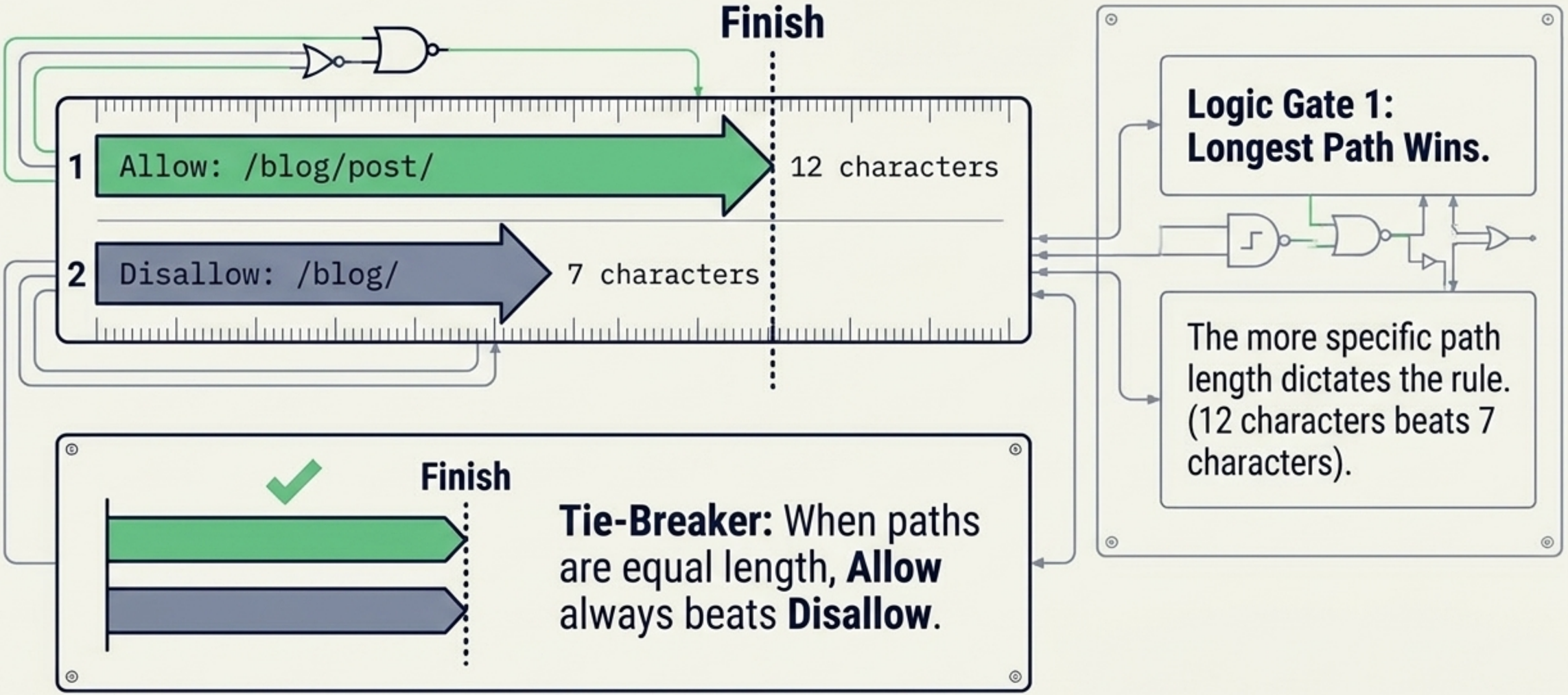
It manages crawl traffic using the **Robots Exclusion Protocol.** A disallowed page can still index if linked externally (appearing without a snippet).

The Golden Rule: Never block JavaScript or CSS.

Why? Googlebot renders pages via headless Chromium. Blocking `.js` or `.css` degrades rendering and triggers "Indexed, though blocked by robots.txt" GSC warnings.

`.js` `.js` `.css`

The Rule of Precedence



2026 AI Bot Exclusion Matrix

Fleet	Model Training	Live Search
Google	Google-Extended (Gemini/Vertex AI)	Googlebot (**Search Ranking**)
OpenAI	GPTBot	OAI-SearchBot (ChatGPT)
Anthropic	ClaudeBot	Claude-SearchBot
Perplexity	[N/A]	PerplexityBot (Indexing/Sourcing)

Critical Warning
Blocking GPTBot does **NOT** remove you from ChatGPT search results.
That requires blocking **OAI-SearchBot** separately.

XML Sitemaps: The VIP List

sitemap.xml

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>https://example.com/</loc>
    <lastmod>2024-10-27T14:30:00+00:00</lastmod>
    <changefreq>daily</changefreq>
    <priority>1.0</priority>
  </url>
  <url>
    <loc>https://example.com/blog/</loc>
    <lastmod>2024-10-25T10:00:00+00:00</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.8</priority>
  </url>
</urlset>
```

Hard Limits

Max 50,000 URLs per file.
Max 50 MB uncompressed.
(Beyond this requires a sitemap index file).

What Google Ignores

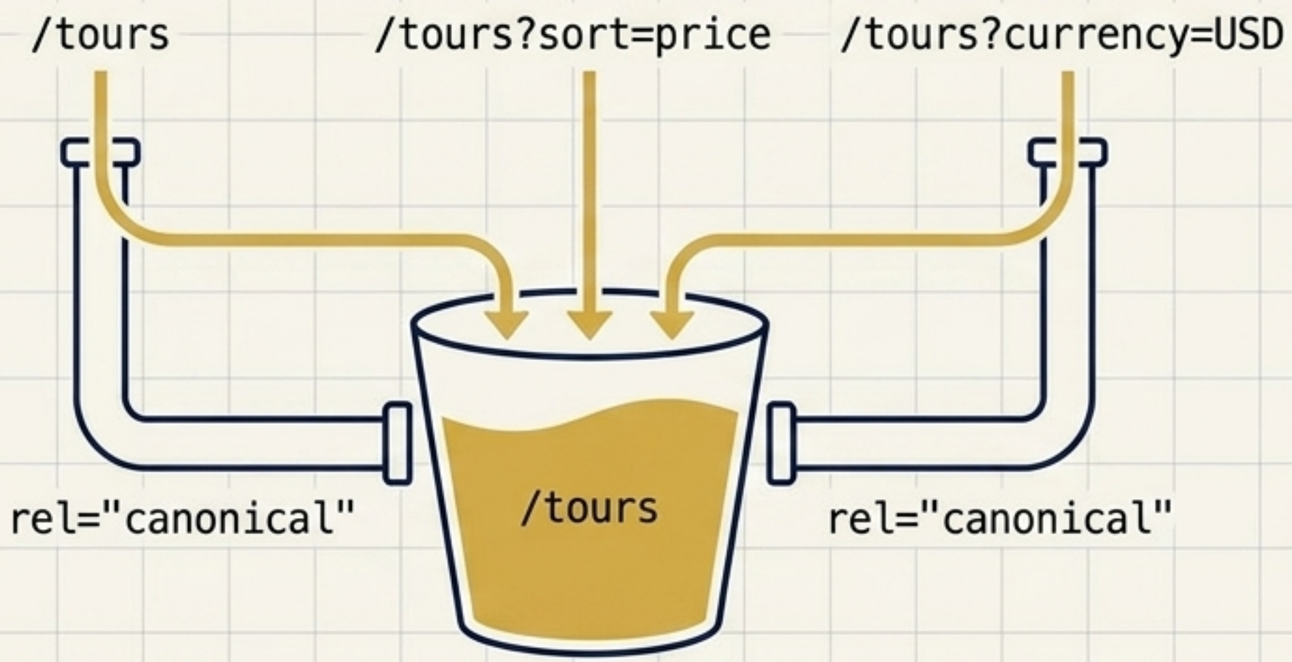
~~<changefreq>~~ and ~~<priority>~~.
(Setting ~~<priority>1.0</priority>~~ provides zero benefit).

What Google Uses

<lastmod>, but only if it verifiably matches actual page modification dates.

Routing Link Equity

Canonical Equity Funnel



Canonical: Consolidates power from duplicate variants.

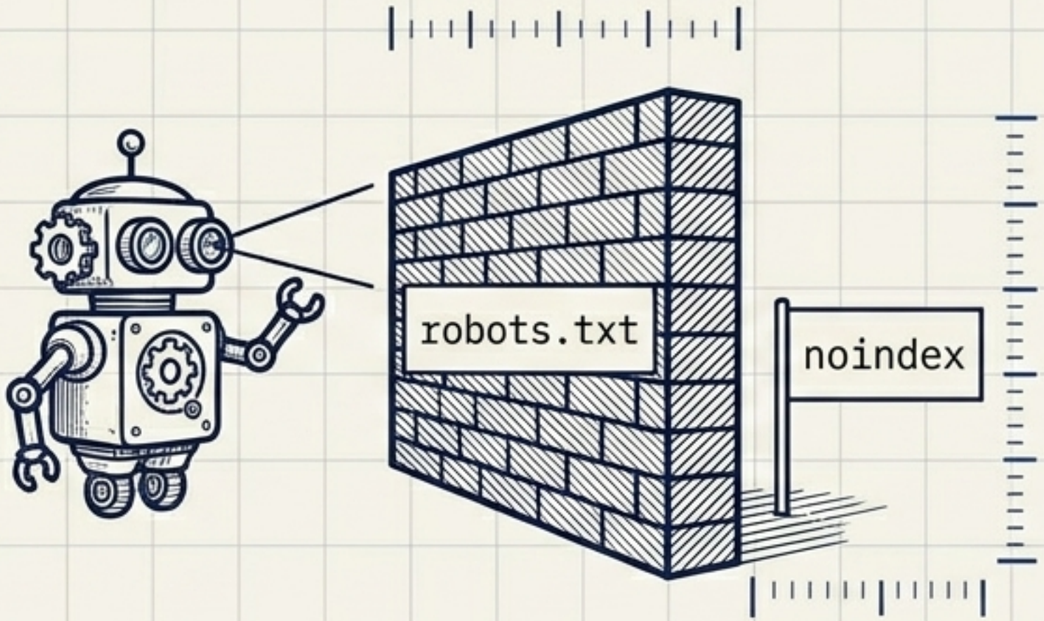
Noindex Drain

Noindex: Discards link equity entirely.

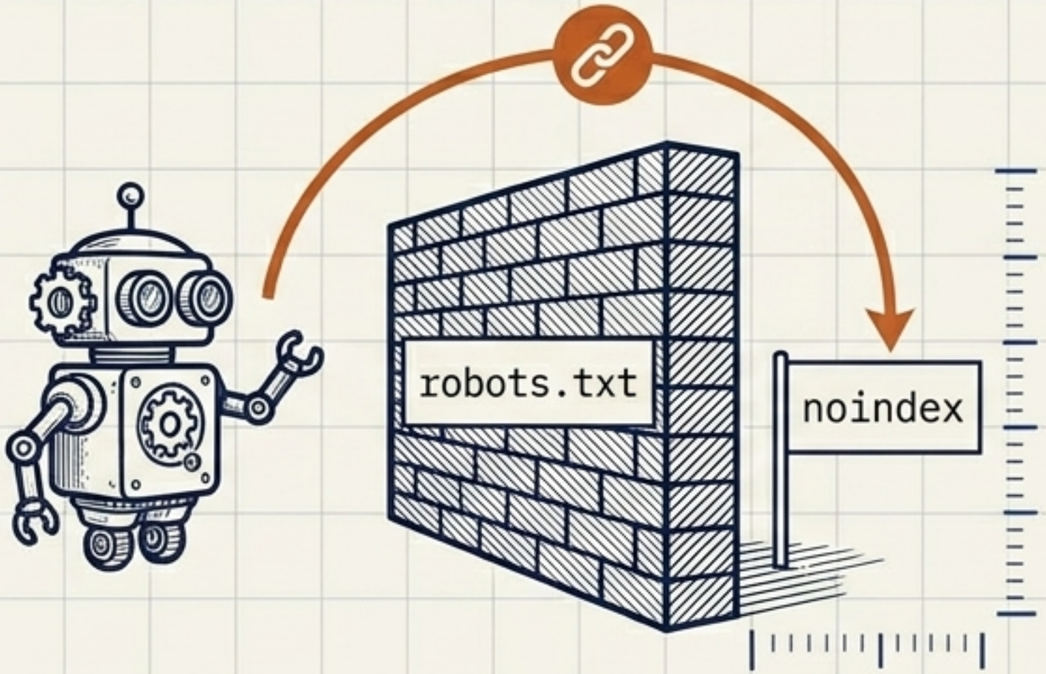
Duplicate Control Diagnostic Matrix

	`Rel=Canonical`	`Noindex`
Removes from Index?	Hint (Google may override)	Directive (Absolute)
Consolidates Link Equity?	Yes ✅	No
Requires Crawlability?	Yes	Yes (Must read the tag)
Ideal OTA Use Case	Filtered listings (`?sort=price`), Print-friendly itineraries.	Internal search results, checkout sequences.

The “Hidden Noindex” Paradox

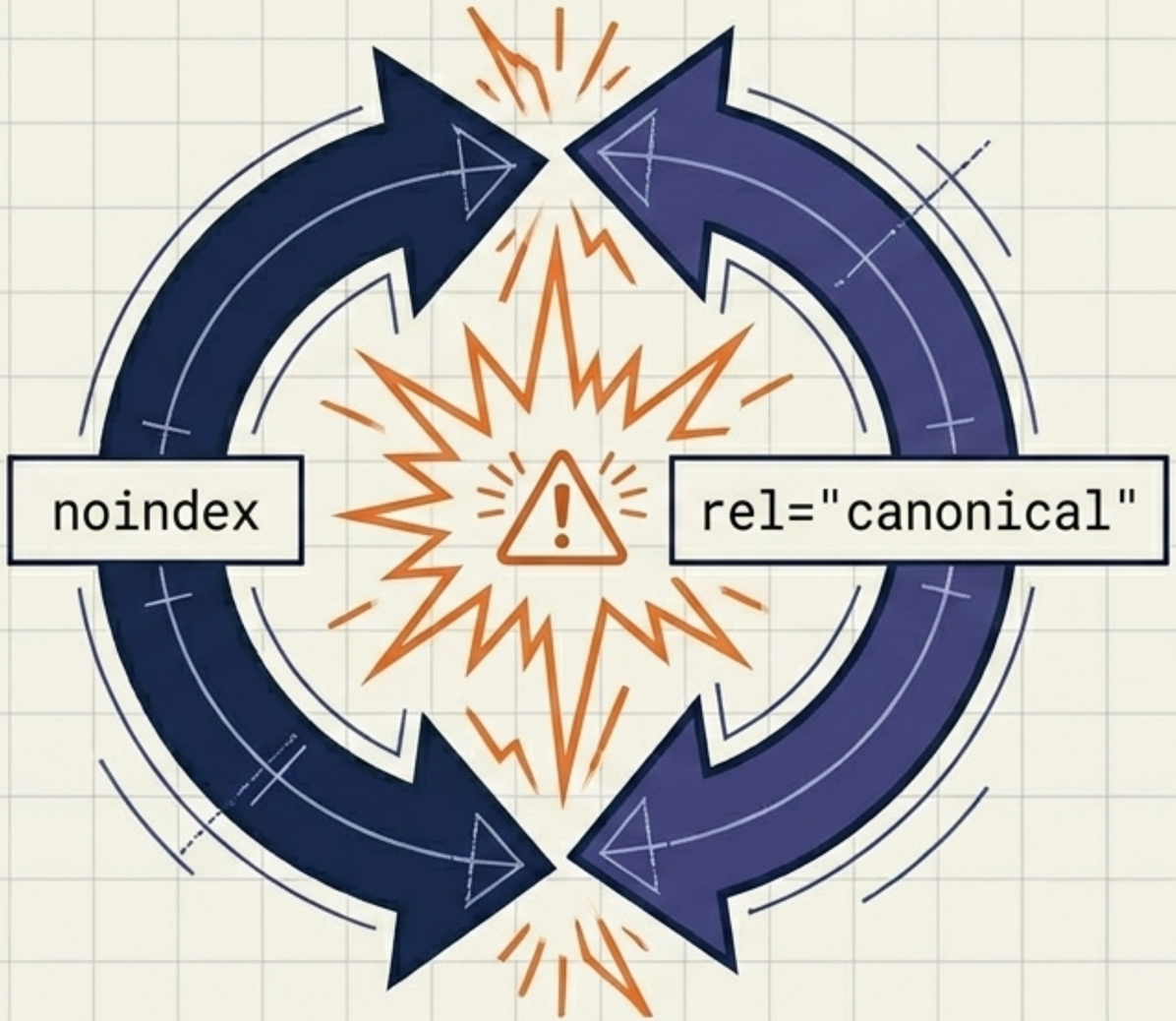


The Conflict: A page blocked by robots.txt is never crawled. If it's never crawled, Google never reads the <meta name="robots" content="noindex"> tag.



The Result: The URL can still be indexed via external link discovery. To permanently exclude a page, it must be crawlable.

The Fatal Combination

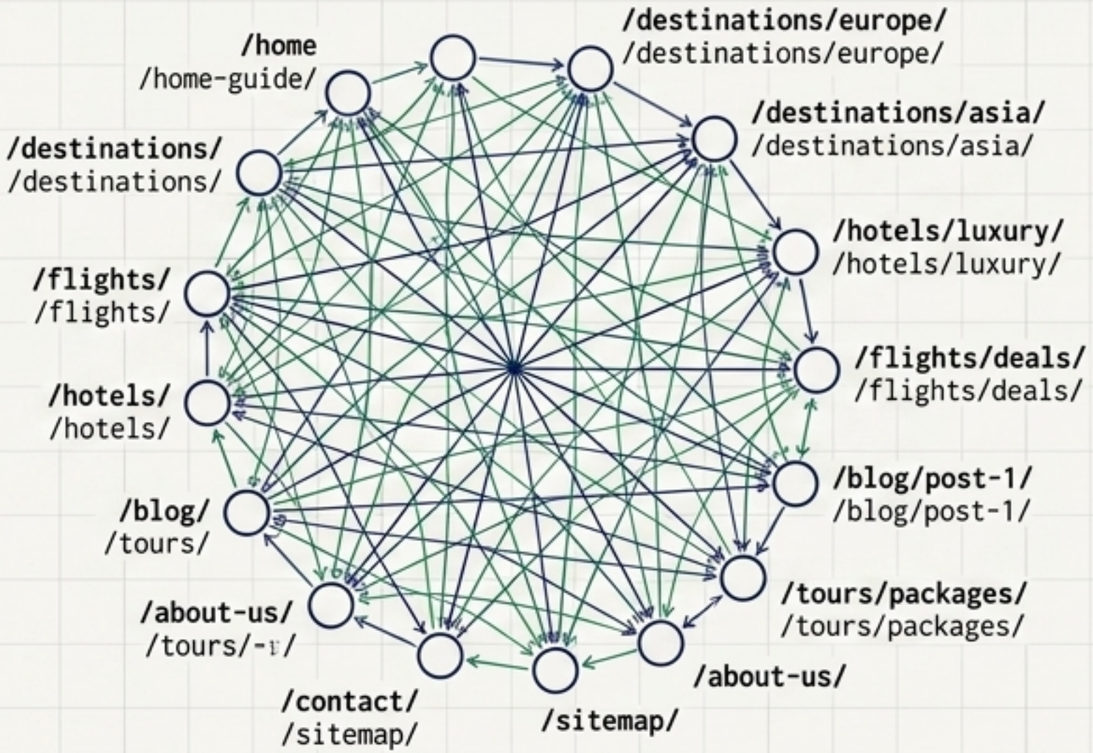


Rule
Never use `noindex` **and** `rel="canonical"` **on the same page.**

The Logic Break
`noindex` declares “Do not include this page in search.”
`rel="canonical"` declares “This page is equivalent to an indexed URL.”

Google’s Response
Contradictory signals force the search engine to guess, risking the de-indexation of the canonical target.

Architecture & The Orphaned Page



Rome Destination Guide
/rome-guide/

The Diagnostic: Internal link gap audits aren't just for link equity—they are crawlability diagnostics.

The Blind Spot: Googlebot navigates via links. Orphaned pages are invisible to the link-following pass and receive minimal crawl budget, regardless of sitemap inclusion.

GSC Page Indexing Triage

Error

Not indexed; problem exists (e.g., 5xx, redirect error).

`Action: Investigate immediately.`

Valid with Warning

Indexed, but issues present.

`Action: Review case-by-case.`

Valid

Indexed normally.

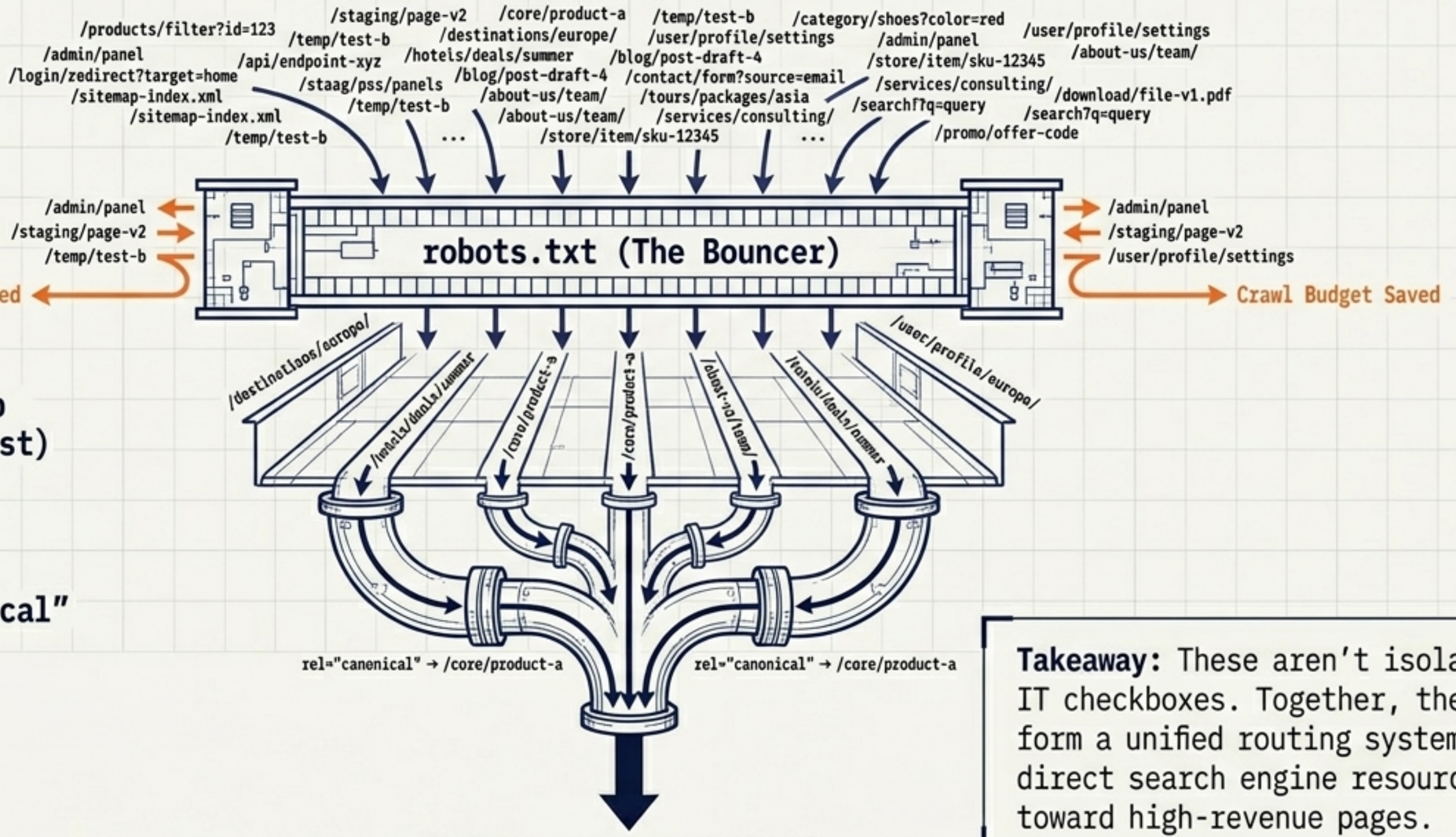
`Action: Monitor`

Excluded

Not indexed; intentional or acceptable deduplication.

`Action: Verify. (Do not treat total Excluded count as a failure metric)`

The Unified SEO Control System



1

2 XML Sitemap (The VIP List)

3 rel="canonical"

Takeaway: These aren't isolated IT checkboxes. Together, they form a unified routing system to direct search engine resources toward high-revenue pages.

The 15-Minute Baseline Triage

Step 1: The Bouncer Check

Open `robots.txt`.

Are `GPTBot` and `OAI-SearchBot` **explicitly** managed?

Use `GSC Settings` to test **parsing**.

Step 2: The VIP List Check

Open `GSC > Sitemaps`.

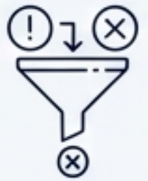
Verify the `URL path` **actually** matches the `submitted file`.

Check the **last-read date**.

Step 3: The Indexing Triage

Open `GSC > Pages`.

Isolate the top `'Error'` **reason** (`Investigate`) and verify the top `'Excluded'` **reason** is intentional.



Next Phase: Full workflow transition to `Screaming Frog` crawl analysis.