

"I Built a 60,000-Page AI Website for \$10: GPTBot Crawled It 30,000+

Metehan Yesilyurt

AI Search & Information Retrieval Researcher · metehan.ai

Published: March 04, 2026 ·

Canonical: <https://metehan.ai/blog/i-built-a-60-000-page-ai-website-for-10-gptbot-crawled-it-30-000-times-in-12-hours/>

A wild experiment.

Why I Built This

Let me be clear upfront: **this website is designed purely as an experiment**. I wanted to observe what real traffic looks like on a large-scale programmatic SEO site and more importantly, how AI crawlers behave in the wild.

This is not a guide on how to build a sustainable business with AI-generated content. If you create programmatic SEO pages only for traffic, one of two things will happen:

1. **Your traffic will tank within weeks** due to deindexing, Google's systems are increasingly good at detecting thin, templated content at scale
2. **You'll see initial traction, then a slow bleed over a few months** as manual reviews or algorithm updates catch up

I've seen this pattern play out repeatedly across the industry. The economics of generating pages are now so cheap that the barrier is essentially zero which is exactly why Google has gotten aggressive about it.

So why build it? Because **the interesting part isn't the SEO. It's the bots**.

I did not expect GPTBot to crawl a brand-new, zero-backlink domain at the scale it did. That was the real discovery.

The Experiment

I built StateGlobe.com,^[1] a Statista-style statistics website covering digital marketing, SEO, content marketing, and web technology across 200 countries. Every single page was generated by AI.

- **60,000 pages** generated in under 30 minutes
- **Total API cost: less than \$10**
- **Model used:** `gpt-4.1-nano` via OpenAI's Chat Completions API
- **Hosted on Cloudflare Workers + D1** (serverless, edge-rendered)

The entire project is open source.

The Tech Stack

Content Generation Pipeline

The pipeline is straightforward:

1. **Taxonomy:** 300 statistical topics × 200 countries = 60,000 unique combinations
2. **Generation:** A Node.js script fires real-time API calls to `gpt-4.1-nano` with controlled concurrency (50 parallel requests) and a token bucket rate limiter
3. **Output:** Each page gets a title, meta description, 5 key statistics, 3 analysis paragraphs, and 2 FAQ items. All as structured JSON
4. **Import:** Results are bulk-imported into Cloudflare D1 (serverless SQLite)

The prompt asks the model to produce 2026 projections based on industry trends. Each response costs fractions of a cent, `gpt-4.1-nano` with `max_tokens: 700` and `response_format: json_object`.

Cloudflare Architecture

- **Cloudflare Workers:** Edge-rendered HTML, no build step, no static files. Every page is assembled on-demand from D1 data
- **Cloudflare D1:** Serverless SQLite storing all 60,000 pages and visit analytics
- **Dynamic OG Images:** Generated on-the-fly as PNGs using `@resvg/resvg-wasm` with the Inter font loaded from a CDN. No pre-generated images, no storage costs
- **Clean URLs:** `/brazil/seo-budget-allocation-statistics` No `.html` extensions, proper 404 headers
- **SEO:** Structured data (Article, FAQPage, BreadcrumbList), XML sitemaps (paginated), canonical URLs, internal linking (same topic across countries, same country across topics)

Total hosting cost: effectively free on Cloudflare's free tier.

What Happened Next: The Bots Arrived

Within minutes of deploying, **GPTBot** started crawling. Hard.

First 12 Hours

- **29,000+ requests from GPTBot alone**
- GPTBot was hitting the site at roughly **1 request per second**, systematically crawling through pages
- OAI-SearchBot and ChatGPT-User also showed up
- GoogleOther appeared with 60+ requests
- Googlebot, AhrefsBot, and PerplexityBot followed

Video oynatıcı yapılandırma hatası

Hata 153

Videoyu YouTube'da izleyin

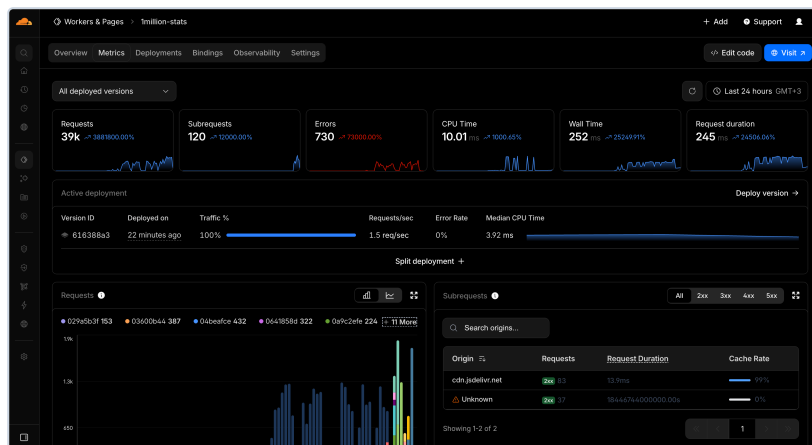
Daha fazla bilgi edinin

Hata 153

3-Hour Snapshot after Server Side Tracking Enabled

Bot	Requests
GPTBot	5,200+
GoogleOther	140+
OAI-SearchBot	94
Googlebot	11
AhrefsBot	7
PerplexityBot	2
ChatGPT-User	1

By the time server-side tracking was fully operational, Cloudflare's own analytics showed **37,000+ total requests** to the worker.



39k requests by almost GPTBot

GPTBot was by far the most aggressive crawler, more active than Googlebot by orders of magnitude.

The Part I Didn't Expect

I've launched plenty of sites before. I expected Googlebot to show up, maybe some SEO tool crawlers. That's normal.

What I did **not** expect was OpenAI's GPTBot hitting a brand-new domain — with zero backlinks, zero social shares, no Search Console submission, at **1 request per second** within minutes of deployment. It found the site through the XML sitemap and just started consuming everything.

This raises serious questions. If GPTBot is this aggressive on fresh domains, how much of the open web is it processing daily? And what does it mean for site owners who haven't explicitly blocked it in `robots.txt`?

For context: Googlebot made 11 requests in the same period that GPTBot made 5,200+. That's a **470x difference** in crawl intensity.

Building Public Analytics

I wanted anyone to see what was happening in real time, so I built a **public analytics dashboard** at stateglobe.com/analytics^[2].

Server-Side Tracking

Initially, I used a client-side beacon (`navigator.sendBeacon`). But bots don't execute JavaScript, so I was missing all bot traffic. The fix was server-side tracking:

- Every request to the Worker records `page_slug`, `user_agent`, `country` (from `cf-ipcountry`), and `is_bot` directly into D1
- Bot detection runs against a pattern list (GPTBot, Googlebot, AhrefsBot, etc.)
- `ctx.waitUntil()` ensures the D1 write completes without blocking the response
- The client-side beacon was removed entirely — one clean tracking path

The Dashboard Shows

- **Human vs. Bot traffic** with separate summary cards (today, this week, all time)
- **Daily traffic chart** (inline SVG line chart, last 30 days, human vs. bot)
- **Top pages, Top bots, Top human user agents**
- **Recent visits** with bot badges, paginated
- Pre-tracking estimates included in totals with clear notes

IP Verification: Catching Spoofed Bots

User agent strings can be spoofed by anyone. A `curl` request with `-A "GPTBot"` would be counted as a real bot visit.

So I implemented **IP verification for OpenAI bots**:

1. Downloaded the official IP ranges from [openai.com/gptbot.json](#)^[3], [searchbot.json](#)^[4], and [chatgpt-user.json](#)^[5]
2. Built a CIDR matching engine directly in the Worker, parses IP ranges into bitmasks for efficient lookup
3. When a request claims to be GPTBot, OAI-SearchBot, or ChatGPT-User, the source IP (from Cloudflare's `cf-connecting-ip` header) is checked against the official ranges
4. **If the IP doesn't match -> classified as human** (potential spoofer)

This means the bot counts on the analytics page are verified — only requests from OpenAI's actual infrastructure count as OpenAI bot traffic.

I verified this works: a `curl` request from Turkey with the GPTBot user agent correctly shows up as a human visit, not a bot.

Hiding Content from AI Crawlers

Here's an interesting twist. I added an "About This Experiment" box on the homepage explaining that the site is AI-generated. But I didn't want GPTBot to read it and potentially use it to discount the content.

The solution: **render it client-side only**.

The HTML source contains an empty `<div>` and a `<script>` tag with the content Base64-encoded. Browsers decode and display it normally. Bots that don't execute JavaScript see nothing — and even bots that parse JS source code can't read Base64 without decoding it.

Verified with `curl` as GPTBot: zero matches for "experiment", "AI Slop", or "Metehan" in the response.

Cost Breakdown

Item	Cost
OpenAI API (gpt-4.1-nano, 60k pages)	~\$10
Cloudflare Workers	\$0 (free tier)
Cloudflare D1	\$0 (free tier)
Domain (stateglobe.com)	~\$10/year
OG image generation	\$0 (on-demand, no storage)
Total	~\$20

The entire 60,000-page website costs about \$20 to create and run.

A Warning About Programmatic SEO

I want to be honest about this: **don't do what I did and expect to rank**.

The economics are seductive. 60,000 pages for \$10, hosted for free. But Google has seen this playbook a thousand times. Here's what typically happens:

1. **Week 1-2:** Pages get indexed, you see impressions in Search Console
2. **Week 3-4:** Traffic trickles in, you feel validated
3. **Month 2-3:** Google's quality systems catch up. Pages start getting deindexed. Traffic drops off a cliff
4. **Month 4+:** Most pages are removed from the index entirely

I've watched this cycle happen to dozens of programmatic SEO projects. The content quality bar keeps rising, and mass-generated pages, even well-structured ones with proper schema, sitemaps, and internal linking — simply don't pass muster anymore.

This site will almost certainly get deindexed. That's not a failure. It's the expected outcome. The experiment was never about ranking. It was about observing the ecosystem.

Figure

Key Takeaways

1. **GPTBot is more aggressive than Googlebot.** On a fresh domain with zero authority, GPTBot made 470x more requests than Googlebot in the first 24 hours. It found and crawled the site almost immediately with no prompting.
2. **Bot traffic dwarfs human traffic.** Bots made up ~98% of all requests. If you're not tracking server-side with user agent analysis, you're blind to what's actually happening on your site.
3. **User agent verification matters.** Without IP verification, anyone can spoof bot traffic. OpenAI publishes their IP ranges, use them.
4. **The barrier to creating massive websites is zero.** 60,000 pages, \$10, 30 minutes. This is simultaneously impressive and terrifying for the health of the open web.
5. **Programmatic SEO for traffic alone is a dead end.** Google will catch it. Build real value or don't bother.
6. **Cloudflare Workers + D1 is a solid stack** for this kind of project. Edge rendering, serverless SQLite, free tier that handles thousands of requests. No servers to manage.

Follow the Experiment

- **Live site:** stateglobe.com^[1]
- **Public analytics** (real-time bot & human traffic): stateglobe.com/analytics^[2]
- **LinkedIn:** [linkedin.com/in/metehanyesilyurt](https://www.linkedin.com/in/metehanyesilyurt)^[6]
- **X (Twitter):** x.com/metehan777^[7]

The site is still live and GPTBot is still crawling. Check the analytics page to see the latest numbers.

This website is intentionally another piece of AI Slop. Created as an experiment to observe how AI crawlers interact with programmatically generated content at scale. If you're thinking about doing this for real SEO: don't. Build something genuinely useful instead.

For tools that surface this kind of insight quickly, see [answer engine optimization tools with near-instant insights](#)^[8].

Generative engine optimization was the actual test behind this build. Scaling pages without generative engine optimization discipline just produces noise; scaling with it produces citation coverage.

REFERENCES & SOURCES

- [1] StateGlobe.com, — <https://stateglobe.com>
- [2] stateglobe.com/analytics — <https://stateglobe.com/analytics>
- [3] openai.com/gptbot.json — <https://openai.com/gptbot.json>
- [4] [searchbot.json](https://openai.com/searchbot.json) — <https://openai.com/searchbot.json>
- [5] [chatgpt-user.json](https://openai.com/chatgpt-user.json) — <https://openai.com/chatgpt-user.json>
- [6] [linkedin.com/in/metehanyesilyurt](https://www.linkedin.com/in/metehanyesilyurt) — <https://www.linkedin.com/in/metehanyesilyurt>
- [7] x.com/metehan777 — <https://x.com/metehan777>
- [8] [answer engine optimization tools with near-instant insights](https://metehan.ai/articles/answer-engine-optimization-tools-near-instant-insights/) — <https://metehan.ai/articles/answer-engine-optimization-tools-near-instant-insights/>