Alvin Kibalama
Digital Marketing Lead · Nutcracker Agency
Operational Playbook 02 · Generative Engine Optimisation

The 6 step playbook to get cited in AI search

A battle-tested framework for re-engineering B2B technical content, capturing citation share in *ChatGPT, Perplexity, and AI Overviews*.

AT A GLANCE

Pages 9 pages
Read time 25 min
Domain Generative Engine Optimisation (GEO)
Execution cadence 60 to 90 Day Sprint
Author Alvin Kibalama
Proven result +52% Citation Lift over 90 Days

alvinkibalama.com
Nutcracker Agency · B2B Growth Architecture
Confidential & Actionable Framework

CORE RETRIEVAL ARCHITECTURE

Executive summary

The structural transformation in search discovery: Traditional search engine optimisation was designed around blue links, keyword density, and domain-level backlink equity. Generative search engines (OpenAI SearchGPT, Perplexity Pro, Google AI Overviews) operate on an entirely different mechanism: *semantic vector retrieval, passage extraction, and multi-source synthesis*.

Ranking on page one of Google no longer guarantees that an LLM will cite your brand when answering a buyer's query. This playbook provides the documented six-step engineering framework to re-structure your digital assets, establish machine-readable authority, and capture dominant citation share.

PREREQUISITES

Pre-flight technical requirements

Figure 2.3: Search Evolution: Traditional 10 Blue Links vs Generative Consensus Citation

The fundamental transition from keyword click-through rankings to verified factual extraction and footnote citations.

Before launching an AI citation optimisation sprint, ensure your organisation possesses the requisite baseline assets. Attempting to optimise for generative retrieval without editorial agility or technical access produces negligible return.

Component	Minimum Threshold	Operational Role
Content Inventory	10+ authoritative articles	High-value topical assets with existing organic indexing and impressions.
Search Console Access	Full property access	Extracting long-tail conversational queries and identifying retrieval gaps.
CMS Velocity	Direct template/code control	Ability to edit HTML tables, schema markup, and root server files within 48h.
Diagnostic Query Log	10–20 buyer prompts	Real commercial queries entered by prospects into ChatGPT and Perplexity.

Figure 2.1: The Six-Step AI Citation Retrieval Pipeline

The sequential technical architecture required to achieve persistent footnote citations across modern LLM answer engines.

STEP 01

Step 1: Restructure content for semantic extraction

Generative models do not read articles linearly. When an LLM executes a Retrieval-Augmented Generation (RAG) query, it chunks content into discrete vector embeddings (typically 250–500 tokens). If an answer to a question is scattered across three prose paragraphs, the model's extraction score drops below the synthesis threshold.

The anatomy of a semantic extraction block

1. **Query-matched header:** Formulate section headings as exact buyer questions (e.g. *"How does vendor attribution calculate CAC payback?"* instead of *"Understanding our methodology"*).
2. **The 45-word direct answer:** Immediately beneath the header, state the direct factual answer in 35 to 50 words. Do not use conversational filler or rhetorical openers. State the definition, the metric, or the operational principle clearly.
3. **Supporting structured data:** Follow the direct answer with a comparison table, bulleted specification list, or quantified benchmark. Models prioritise structured tabular data over unstructured prose during synthesis.

Figure 2.2: How Large Language Models Select, Weight, and Cite Sources

Multi-source vector retrieval, cross-source fact verification, and synthesized footnote citation attribution.

STEP 02

Step 2: Build topical authority with comparison data

Over 40% of conversational B2B search queries involve comparative evaluation (e.g., *"[Product A] vs [Product B]"*, *"Best alternatives to [Enterprise Vendor]"*). AI engines prioritise sources that provide neutral, structured comparison matrices over vendor-biased promotional claims.

Structuring comparison tables for machine extraction

When publishing comparison content, use clean semantic HTML (`<table>`, `<thead>`, `<tbody>`). Never render comparison matrices inside JavaScript sliders, images, or client-side accordions. State trade-offs objectively: acknowledge limitations of your own platform and identify specific use cases where a competitor is better suited. Models penalise unilateral bias.

STEP 03

Step 3: Make your authority legible via llms.txt

The `llms.txt` standard is an emerging specification designed to provide AI crawlers with a curated Markdown index of a website's core knowledge base, eliminating HTML boilerplate, JavaScript rendering overhead, and tracking scripts.

Production llms.txt implementation structure

```
# Alvin Kibalama: Generative Search Engine Specification
> Enterprise B2B digital marketing, channel attribution, and AI retrieval architecture.

## Canonical Frameworks
- [The B2B Buyer Shortlist Playbook](https://alvinkibalama.com/playbooks/b2b-buyer-shortlist-pl
- [AI Citation Optimisation Playbook](https://alvinkibalama.com/playbooks/ai-citation-optimisat
- [SEO vs PPC Decision Framework](https://alvinkibalama.com/seo-vs-ppc-b2b-decision-framework/)

## Verified Operational Rules
- Content Chunking: Maximum 350 tokens per semantic answer block.
- Heading Case: Sentence case applied across all technical documentation.
- Attribution Standard: Triangulated measurement (Software, Self-reported, Econometric).
```

STEP 04

Step 4: Make your commercial pages citable

Standard sales and pricing pages are notoriously difficult for AI search engines to parse because they are heavily gated, laden with vague marketing adjectives (*"revolutionary platform"*, *"seamless synergy"*), or hide pricing behind contact forms. When an AI agent attempts to evaluate your product for a buyer, it skips pages that lack concrete factual parameters.

The above-the-fold factual brief

On every core service and product page, place a clean 3-row summary card directly beneath the hero header:

- **Core Architecture:** Exactly what the technology or service does in plain technical language.
- **Deployment Requirements:** Required integrations (CRM, SSO, cloud infrastructure) and onboarding timeframe.
- **Commercial Model:** Pricing tiers, contractual minimums, or implementation fee structures.

STEP 05

Step 5: Eliminate technical crawler friction

AI search crawlers (such as GPTBot, PerplexityBot, and Google-Extended) enforce aggressive timeout thresholds when scraping pages during real-time synthesis queries. If your server response time (TTFB) exceeds 1.2 seconds, or if content requires client-side JavaScript hydration, the crawler falls back to a competitor.

Technical crawler hygiene checklist

- **Server-Side Rendering (SSR):** Ensure all informational content is delivered in raw server-rendered HTML. Verify by fetching your URL via `curl -s https://yourdomain.com | grep "your core content"`.
- **JSON-LD Schema Markup:** Deploy detailed `TechArticle`, `SoftwareApplication`, and `organisation` schemas with `sameAs` arrays linking to Wikipedia, Wikidata, and verified social profiles.
- **HTTP Status Integrity:** Eliminate 301 redirect chains. Ensure all internal links resolve directly to the canonical 200 HTTP destination.

Production JSON-LD schema example

```
{
  "@context": "https://schema.org",
  "@type": "TechArticle",
  "headline": "The 6 step playbook to get cited in AI search",
  "author": {
    "@type": "Person",
    "name": "Alvin Kibalama",
    "url": "https://alvinkibalama.com/about/",
    "sameAs": [
      "https://www.linkedin.com/in/alvinkibalama/",
      "https://x.com/alvinkibalama"
    ]
  },
  "publisher": {
    "@type": "Person",
    "name": "Alvin Kibalama"
  },
  "inLanguage": "en-GB"
}
```

STEP 06

Step 6: Optimise for autonomous AI agents

AI agents do not just browse websites; they complete multi-step tasks on behalf of human researchers. When an agent visits your site to verify compliance or feature availability, it looks for standardized paths and clear exit points.

Auditing and tracking share of model voice (SoMV)

Calculate your share of model voice across target evaluation prompts on a monthly schedule:

```
SoMV = (Total Brand Citations in Synthesis Runs / Total Industry Citations Observed) × 100
```

Run 20 standardized prompt variations in incognito environments across ChatGPT-4o Search, Perplexity Pro, and Gemini 1.5 Pro. Log whether your brand is: (a) Cited as primary source, (b) Mentioned neutrally, or (c) Omitted completely.

EXECUTION CADENCE

60-day sprint implementation checklist

- **Week 01:** Identify top 15 existing articles driving organic impressions. Extract customer questions from GSC query reports.
- **Week 02–03:** Restructure articles with 45-word direct answer blocks beneath query-formatted H2 headers.
- **Week 04:** Deploy HTML comparison tables addressing top 3 category competitors with objective pros, cons, and limits.
- **Week 05:** Author and publish `/llms.txt` and `/llms-full.txt` at domain root. Validate robots.txt permits AI crawlers.
- **Week 06:** Implement `TechArticle` and `organisation` JSON-LD schema with complete entity linkage.
- **Week 07:** Benchmark server response velocity (TTFB <1.2s) and audit raw HTML responses using command-line curl.
- **Week 08:** Conduct initial share of model voice baseline audit across 20 primary commercial prompts.

NUTCRACKER AGENCY · ADVISORY & EXECUTION

Put this playbook into commercial context.

Alvin Kibalama is a digital marketing consultant and growth architect based in the United Kingdom. He advises B2B enterprise software and professional services firms on organic search engineering, paid media allocation, and AI retrieval optimisation.

alvinkibalama.com

Enquiries: alvin.kibalama@nutcrackeragency.com