



SCAN TO
READ ONLINE

← PYC • GUIDES

GUIDE / GENERATIVE-ENGINE-OPTIMISATION

Getting *cited*, not just ranked.

Generative engine optimisation — being named as the source inside an AI answer, not just ranked beneath it. What actually decides citation, grounded in Google's own guidance, the founding research, and real field data — with runnable code.

By Phil Yarrow • 17 min read • updated 2026-07-16

In one minute

THE WHOLE ARGUMENT, UP FRONT

- 01 Generative engine optimisation (GEO) is getting your page named as the source in an AI answer — an AI Overview, ChatGPT, Perplexity — not just ranked in the blue links. In the AI era that is the citation that compounds, because the answer often resolves the query without a click.
- 02 It is earned the same way authority always was. Google is explicit: there is no special AI markup, file or trick, and no schema that unlocks AI Overviews — a page just has to be indexed, snippet-eligible, genuinely helpful and credible. GEO is topical authority, pointed at extraction.
- 03 Five levers decide whether you are cited: retrievability (you cannot be quoted if you are not retrieved), passage self-containment (the model lifts a passage, not a page), information gain (it names the source that adds what the others left out), entity clarity (it must resolve who you are to attribute a claim), and extractability (structure it so the answer is cheap to lift).
- 04 The research backs the boring answer. The GEO paper (KDD 2024) found the methods that lift a source's visibility in generative answers by up to ~40% are content-quality signals — citing sources, adding quotations and statistics — not keyword tricks. Fabricated ones fail, because the model and the reader both check.
- 05 Measure it honestly. Citation is still hard to see, so read it through the decoupling — impressions rising while clicks compress — and through the surviving click: authority keeps the click AI answers take from everyone else. In our field data, the account whose position climbed the most (page five to page two) saw its click-through rise even as the era compressed it elsewhere — earning authority beat the decoupling.

Ranked gets you shown. Cited gets you quoted.

For twenty years the goal of search was a **position** — be the blue link people click. Answer engines changed the goal. Now a query can be met by a synthesised answer that quotes a handful of sources and names them, and the win is being one of those **citations**, not just ranking somewhere in the list underneath. That is generative engine optimisation.

It is tempting to treat this as a new dark art with its own tricks. It is not, and Google has said so as plainly as it ever says anything: there is *no* special markup, no AI text file, no schema that unlocks AI Overviews, and no optimisation beyond meeting the normal technical requirements and being eligible to show with a snippet. GEO is not a hack bolted onto SEO. It is [topical authority](#) pointed at one new thing: being extracted.

So this guide is the opposite of a shortcut. It is the mechanics of why an engine cites one source and not the nine that say the same thing — and the honest, testable work that makes you the one.

Why it matters now

Because the click is decoupling from the impression. As answer engines resolve more queries on the results page, the same visibility earns fewer clicks — impressions up, click-through down. We measured it across a [16-month study of five live businesses](#): the one account that showed a clean AI-era decoupling had impressions up around 230% while clicks barely moved.

The response is not to mourn the click. It is to change what you compete for. Two things still hold their value in a decoupled world: being **cited** — named as the source at the moment of the answer — and ranking high enough to keep the click for the queries that still click. And authority buys both. In the same study, the account whose position climbed the most saw its click-through *rise*

while the era compressed everyone else's, and Google itself reports that the clicks AI experiences do send tend to be higher quality. Cited, or the survivor of the click — either way the currency is authority, not traffic volume.

The anatomy of a citation

An answer engine generally works in this order: it **retrieves** candidate passages (largely from the search index), **selects** the passages that answer the query, **grounds** its sentences on them, and **attributes** — names the sources it grounded on. You are cited when your passage is what a claim is built on.

And it rarely answers your query as you wrote it. Engines like AI Overviews fan a single query out into a set of **sub-queries** and retrieve for each — so your passage gets scored against questions you never targeted. That is the mechanical reason a passage has to answer *on its own*: it may be lifted to settle a sub-query the rest of your page never mentions.

The engines differ under the hood — AI Overviews and AI Mode retrieve from Google's index with that query fan-out; ChatGPT leans on Bing's; Perplexity runs its own index and cites heavily — but they share this retrieve, select, ground, attribute shape. Optimise for the shape, not one engine's quirks. Eleven terms, defined so nothing rides on jargon:

Generative engine optimisation (GEO)

Optimising so a generative answer engine names your page as a source, not just so a search engine ranks it. The unit of success moves from a blue-link position to a citation inside a synthesised answer.

Answer engine

A system that synthesises a direct answer from retrieved sources and cites some of them — Google's AI Overviews and AI Mode, ChatGPT with search, Perplexity, Gemini. It retrieves, then generates, then attributes.

Retrieval

The step where the engine pulls candidate passages to answer from — largely from the search index and the pages already eligible to rank. Retrieval is the gate: a page that cannot be retrieved cannot be cited, whatever else is true of it.

Passage

The chunk the engine actually lifts and grounds an answer on — a paragraph or section, not the whole page. Citation is won or lost at the passage level, which is why a page can rank yet never be quoted.

Query fan-out

A single query decomposed into a set of related sub-queries, each retrieved for separately and then recombined into one answer. It is the mechanical reason a passage must stand alone: it can be lifted to settle a sub-query your page was never written for.

Grounding & attribution

Grounding is the engine tying a sentence in its answer to a retrieved passage; attribution is naming the source that passage came from. You are cited when your passage is what a claim is grounded on.

Information gain

The information a page adds beyond what already ranks — the stat, the caveat, the case the other results omitted. Google's Information Gain thinking rewards it, and it is the single strongest reason an engine cites you specifically rather than the ten pages that agree.

Entity clarity

How unambiguously a machine can resolve who you are and that you are credible on the subject. To attribute a claim to you, the engine has to know you exist as a trusted entity — so entity consolidation is a citation prerequisite, not a nicety.

Extractability

How cheaply an engine can lift a self-contained answer from your page — a direct answer up front, definitions, lists, tables, valid structured data, clean HTML. The lower the cost of extraction, the more liftable your passage.

Citation share

Of the AI answers for the queries you target, the share that name you as a source. The GEO metric — the one that replaces rank as the thing you are actually competing for in a generative answer.

The great decoupling

The AI-era pattern where impressions rise while clicks and click-through fall, because answers resolve queries on the results page. It is why “shown” is worth less than it was, and why “cited, and ranked high enough to keep the click” is the durable play.

THE FUNNEL

The five gates to a citation

A passage of yours passes five gates to end up cited. Each is opened by a different lever, and a failure at any gate ends the run — you can be perfectly authoritative and still never quoted if your answer is not liftable, or perfectly written and never retrieved.

- 01 **Retrieved** TOPICAL AUTHORITY
INDEXED, SNIPPET-ELIGIBLE, IN THE CANDIDATE SET AN ANSWER IS BUILT FROM.
- 02 **Passage selected** SELF-CONTAINMENT
A BLOCK THAT ANSWERS THE QUERY ON ITS OWN, NOT THE WHOLE PAGE.
- 03 **Grounded on** INFORMATION GAIN
A DISCRETE, VERIFIABLE CLAIM WORTH BUILDING THE ANSWER ON.
- 04 **Attributed** ENTITY CLARITY
A SOURCE THE ENGINE CAN RESOLVE AND TRUST ON THE SUBJECT.
- 05 **Cited** EXTRACTABILITY
STRUCTURED SO LIFTING THE ANSWER IS CHEAP.

Copper = the lever that opens each gate. The first two are where most sites fail: they never get retrieved (no authority), or they rank but have no self-contained passage to lift.

The levers, in order

The five gates become five moves, in sequence, plus the discipline of feeding the result back. The order matters: there is no point structuring a passage the engine will never retrieve.

01

Earn retrieval first

You cannot be cited if you are not retrieved, and retrieval is gated by the same thing rankings are: topical authority. Complete, credible coverage of the subject gets your pages indexed, snippet-eligible and into the candidate set an answer is built from. No authority, no retrieval, no citation — everything below assumes this is in place.

02

Answer the query in a self-contained passage

The engine lifts a passage, not a page. Lead each section with the answer to the question it targets, stated so it stands on its own without the surrounding context. A page that buries its answer, or only makes sense read whole, ranks but does not get quoted.

03

Add the information gain

Give the engine a reason to cite you rather than the ten pages that say the same thing: the real number, the honest caveat, the case the others left out. The GEO research is blunt about this — citations, quotations and statistics are what lifted visibility, because they are the discrete, verifiable claims an answer can ground on.

04

Make the source unmistakable

An engine can only attribute a claim to a source it can resolve. Consolidate your entity — a consistent identity, sameAs links to authoritative profiles, a clear source context — so the machine knows who you are and that you are credible on this subject. Ambiguity is uncitable.

05

Structure for extraction

Lower the cost of lifting your answer: a direct answer first, definitions, lists and tables where they fit, valid structured data, clean and fast HTML. You are not adding magic AI markup — Google says there is none — you are making a genuinely good answer cheap for a machine to read and quote.

06

Measure the surviving signal, and feed it back

Citation is hard to see directly, so watch the decoupling (impressions vs clicks), branded-query lift, and the click authority keeps when others lose it — then feed what you learn back into the coverage. GEO is not a one-off pass; it is topical authority maintained as the answer surfaces change.

EVIDENCE • RUNNABLE

Test it yourself

Two of the levers are measurable with the same open-source primitives an answer engine runs on — so you can check, not guess, whether a page is liftable and whether it adds anything worth citing.

1· Is there a passage worth lifting?

An engine quotes a passage, not a page. Chunk the page, score each chunk against the query, and see whether anything self-answers. If nothing clears the bar, you may rank and still never be quoted. Uses [sentence-transformers](https://sbert.net/)

(<https://sbert.net/>).

```
# pip install sentence-transformers
from sentence_transformers import SentenceTransformer
model = SentenceTransformer("all-MiniLM-L6-v2")

# An engine lifts a PASSAGE, not a page — so score the page passage by passage.
query = "can misted double glazing be repaired without replacing the frame"
page = open("your-page.txt").read()
passages = [p.strip() for p in page.split("\n\n") if p.strip()] # paragraph-level chunks

q = model.encode(query, normalize_embeddings=True)
scores = model.encode(passages, normalize_embeddings=True) @ q # cosine per passage
best = int(scores.argmax())

print(f"best passage ({scores[best]:.2f}):", passages[best][:160])

# If no passage clears ~0.5, the page may rank but has nothing self-contained to lift.
# Rewrite the target section to answer the query on its own, up front.
```

2· Do you add anything worth citing?

The engine grounds on the source that added the entity, the number, the caveat the others omitted. Measure what your page carries that a currently-cited answer does not — the information gain that is your reason to be quoted. Uses [spaCy](https://spacy.io/) (<https://spacy.io/>).

```
# pip install spacy && python -m spacy download en_core_web_sm
import spacy, re
nlp = spacy.load("en_core_web_sm")

def claims(text):
    # the discrete, citable things: named entities + any sentence carrying a statistic
    ents = {e.text.lower() for e in nlp(text).ents}
    stats = {s.strip() for s in re.split(r"(?<[.!?]) ", text) if re.search(r"\d", s)}
    return ents, stats

mine, my_stats = claims(open("your-page.txt").read())
theirs, _ = claims(open("top-answer.txt").read()) # a page already cited for the query

print("information gain – entities they lack:", sorted(mine - theirs))
print("citable statistics on your page      :", len(my_stats))
# What you have that the cited answer does not is your reason to be grounded on.
# Zero novel entities and zero statistics = retrievable, not citable.
```

3·Emit the answer as structured data

Extractability is lowering the cost of lifting a clean, attributable answer – not adding magic markup. Emit the Q&A from the same facts, tied to your entity.

```
import json

# Extractability: emit the answer AS structured data from the same facts, so the
# engine can lift a clean, attributable claim — not scrape it out of prose. An
# FAQPage is the idiomatic container (and the one eligible for rich results).
def faq_schema(question, answer, author_id):
    return json.dumps({
        "@context": "https://schema.org",
        "@type": "FAQPage",
        "mainEntity": [{
            "@type": "Question",
            "name": question,
            "acceptedAnswer": {"@type": "Answer", "text": answer,
                                "author": {"@id": author_id}},
        }],
    }, indent=2)

# Structured data does not "unlock" AI answers (Google is clear it is not required) —
# it lowers the cost of reading a self-contained, sourced answer. That is the point.
```

What this proves — and what it doesn't

These measure the concepts citation is built on — passage relevance, information gain, extractable structure — with the same primitives (embeddings, entity extraction, schema) answer engines use. They do *not* reverse-engineer any engine's selection, and nothing here claims to. Google is explicit that no special optimisation exists; these give you an honest read on whether your page is retrievable, liftable and worth citing — the part you actually control.

How to measure it

Citation is still hard to see — the tooling lags the shift. Read it through three proxies, in order of how well they track it:

-
- 01 **The decoupling.** In Search Console, impressions climbing while clicks flatten is the fingerprint of answers resolving your queries. Watch the gap, per query group — it is the clearest signal you are being surfaced in an answer surface.
-
- 02 **Entity and branded-query lift.** People who see you cited search for you afterwards. A rise in branded and entity queries is citation converting into demand — and it reinforces the entity signals that get you cited again.
-
- 03 **Manual citation checks.** Run your target queries in the AI surfaces — AI Overviews, ChatGPT, Perplexity — and record who is named. Crude and unscalable, but it is ground truth, and it tells you which passages of yours actually get lifted.
-
- 04 **The surviving click.** Then the authority metrics — position, non-brand share, and the click-through that holds when others' falls. The [audit](#) reads these with real data; they are the lagging confirmation that the coverage is working.
-

Where it goes wrong

The six ways GEO gets chased in the wrong direction:

Chasing AIO “hacks”

Hunting for a secret markup, file or prompt that unlocks AI Overviews. Google states plainly there is none — no special schema, no AI text file, no trick. The play is foundational quality, authority and extractability, and anyone selling a hack is selling nothing.

Writing for rank, not extraction

A page that ranks but has no self-contained passage answering the query never gets quoted. Ranking is necessary and no longer sufficient; if the answer is buried or only makes sense whole, the engine retrieves you and lifts someone else.

No information gain

Publishing the same answer as everyone else. That gets you into the candidate set and cited by no one — the engine grounds on the source that added the number, the caveat or the case the others omitted. Sameness is retrievable but not citable.

Entity ambiguity

Leaving the machine unable to work out who you are. A claim cannot be attributed to a source the engine cannot resolve or does not trust on the subject, so an unclear entity is uncitable no matter how good the passage.

Optimising for a click count AI is compressing

Judging the programme on raw traffic while answer engines quietly resolve more queries without a click. The durable targets are citation share and the surviving, higher-intent click — not a visit count the era is deflating anyway.

Fabricating the citations and statistics

The GEO lever is real citations, quotations and statistics. Inventing them fails twice: the reader who follows the source catches it, and the credibility signal an engine (and E-E-A-T) rewards is precisely the one you have faked. Cite real sources or none.

The honest limits

GEO is a young field moving fast, and a guide that pretends otherwise is selling certainty it does not have. Four boundaries worth naming:

01 The internals of AI Overviews and the other answer engines are not public. The levers here are grounded in Google's own guidance, the retrieval literature and the GEO research — not reverse-engineered mechanics. Treat them as well-founded principles, not a leaked algorithm.

02 It is correlation, not a control group. Our field data shows the decoupling and the authority-keeps-the-click pattern; it does not isolate citation as a single cause. The [study](#) is where the honest, hedged version of the data lives.

03 Measurement is immature. There is no clean “citation share” report yet; the proxies above are the best available and will be replaced as the platforms expose more.

04 It changes monthly. Surfaces, formats and behaviours shift faster than any tactic. Which is exactly why the durable play is authority and genuine information gain — the things that survive every reshuffle — not a format-of-the-month trick.

Questions

Is GEO actually different from SEO?

It is SEO pointed at extraction, not a separate discipline. Google is explicit that there are no additional requirements or special optimisations to appear in AI features — the foundational work (indexable, helpful, credible, well-structured) is what makes you both rankable and citable. What changes is the target you optimise toward: a citation inside an answer rather than only a position beneath it.

Do I need special schema or an AI file for AI Overviews?

No. Google states directly that there is no special schema.org markup, no new machine-readable or AI text file, and no special optimisation needed to appear in AI Overviews or AI Mode — a page just has to meet the normal technical requirements and be eligible to show with a snippet. Structured data still helps a machine read you; it is not a secret unlock.

How do I know if I am being cited?

Imperfectly, for now — the tooling lags. Read it three ways: the decoupling in Search Console (impressions climbing while clicks flatten is the fingerprint of answers resolving your queries), branded and entity-query lift (people searching you after seeing you cited), and manual checks — run your target queries in the AI surfaces and see who is named. Treat it as a trend to watch, not a number to report to two decimals.

Does being cited but not clicked still help?

Yes. A citation is brand and entity exposure at the moment of the answer, it reinforces the entity signals that get you cited again, and Google reports that the clicks that do come from AI experiences are higher quality — users spend longer. Authority also keeps the click others lose: in our field data the account whose position climbed the most saw its click-through rise even as the era compressed it for the rest.

Will AI answers kill organic traffic?

They compress it — the decoupling is real, and some informational queries will resolve without a click. But authority is the hedge, not the casualty: rank high enough and cover the subject completely enough, and you keep the click for the queries that still click and get named for the ones that do not. The businesses that lose are the ones whose traffic was thin to begin with.

So what actually moves the needle?

The unglamorous answer the evidence keeps giving: genuine topical authority so you are retrieved, self-contained passages so you are liftable, real information gain — citations, quotations, statistics — so you are the one grounded on, and a clear entity so you can be attributed. Not tricks. The GEO research measured up to a ~40% visibility lift from exactly those content-quality signals.

Further reading

Google's own guidance, the founding GEO paper, the retrieval literature and the open-source tooling. Real sources, checked, not a citation wall.

POLICY Google Search Central — AI features and your website ↗ (<https://developers.google.com/search/docs/appearance/ai-features>)

Google's own line: no special markup or optimisation for AI Overviews; a page must be indexed and snippet-eligible. The retrievability gate, from the source.

POLICY Google — Guide to optimising for generative AI features ↗ (<https://developers.google.com/search/docs/fundamentals/ai-optimization-guide>)

Google's guidance on performing well in AI experiences: foundational, people-first content, not new tricks.

POLICY Google Search Central Blog — Top ways to succeed in AI Search (2025) ↗ (<https://developers.google.com/search/blog/2025/05/succeeding-in-ai-search>)

The practical version, and the note that clicks from AI experiences tend to be higher quality.

RESEARCH GEO: Generative Engine Optimization — Aggarwal et al., KDD 2024 (arXiv:2311.09735) ↗ (<https://arxiv.org/abs/2311.09735>)

The founding academic paper: content-quality signals — citing sources, adding quotations and statistics — lifted visibility in generative answers by up to ~40%, varying by domain.

RESEARCH Retrieval-Augmented Generation — Lewis et al., 2020 (arXiv:2005.11401) ↗ (<https://arxiv.org/abs/2005.11401>)

The retrieve-then-generate architecture answer engines are built on — why retrieval gates everything downstream.

RESEARCH Google's Information Gain patent, explained — Search Engine Journal ↗ (<https://www.searchenginejournal.com/googles-information-gain-patent-for-ranking-web-pages/524464/>)

A Google patent describing higher scores for pages that add information beyond what already ranks — the idea behind being cited rather than merely retrieved (a patent, not a confirmed live signal).

TOOL sentence-transformers (SBERT) ↗ (<https://sbert.net/>)

Open-source semantic similarity — score passage-level relevance yourself.

TOOL

spaCy — industrial-strength NLP ↗ (<https://spacy.io/>)

Open-source entity extraction — measure the information gain your page carries over the cited answer.

TOOL

Schema Markup Validator ↗ (<https://validator.schema.org/>)

Keep structured data valid so a machine can read a clean, attributable answer.

The proof, and where to go next

GEO is not a separate programme from the rest of the method — it is what topical authority earns you in the AI era. The strategy, the pipeline that ships it, and the data are all here:

Topical authority — what earns the citation

[READ →](#)

The coverage that gets you retrieved and grounded on

The compounding curve — the decoupling, in real data

[READ →](#)

16 months, five businesses, the AI-era click compression

Signage & personalisation e-commerce

£11,244

organic revenue · converts at 2× its traffic weight

Prestige-marque automotive specialist

5+ yrs

retained · 68% of traffic organic · #1 local for the marque

Domestic glazing & window repair

~5×

organic clicks in the recent quarter, YoY · page 5 → page 2

Auto locksmith & vehicle key replacement

505

qualified leads from organic · 77% of new customers

Compliance & quality-management software

Near-zero → p2

category head-term visibility built over 16 months



Cited, not just ranked.

What decides whether an AI answer names you as the source — retrievability, self-contained passages, information gain, entity clarity, extractability. To build it into your own site, book a teardown at pyc.agency.

PYC • GENERATIVE ENGINE OPTIMISATION — GETTING CITED, NOT JUST RANKED • VOL. 26 • JULY 2026
SCAN, OR VISIT [PYC.AGENCY/GUIDES/GENERATIVE-ENGINE-OPTIMIZATION](https://pyc.agency/guides/generative-engine-optimization) • [INFO@PHILYARROW.CO.UK](mailto:info@philyarrow.co.uk)