Prove me wrong

Tracking brand visibility inside most LLM chatbots is, for now, experimental data. It can be interesting telemetry, but it is not yet decision-grade.

By James Crawford
Managing Director of PR Agency One

PR AgencyOne

Why did I write this report?

Generative Engine Optimisation (GEO) is no longer a thought experiment, it is here.

Every major platform now uses some form of generative AI to answer questions, summarise results, or recommend brands. For communicators, that means visibility is shifting from search results pages to conversational interfaces.

The idea is straightforward: if people get answers from AI assistants such as ChatGPT, Google's AI Overviews, or Perplexity, then brands need to understand how, and whether, they appear in those answers.

But the data so far is noisy. Agencies, SEO tools and analytics vendors have poured real effort into testing GEO. Many have been brave enough to publish results publicly, despite limited benchmarks and rapidly changing technology. That transparency deserves credit.

Yet much of what is being shared mixes real experimentation with fragile methods and early assumptions. Some experiments rely on technical interfaces (known as APIs) that do not match what real users see on screen.

Others collapse billions of possible user prompts into a single percentage score and call it "share of model".

This report separates what is measurable from what is guesswork. It recognises that GEO is a real and important discipline, but it also challenges some of the over-confident claims surrounding it.

Above all, it tests one clear proposition:

Tracking brand visibility inside most LLM chatbots is, for now, experimental data, it can be interesting telemetry, but not yet decision-grade data.

Prove us wrong.

PR AgencyOne

Foreword from Stephen Waddington

Generative Engine Optimisation is real, but don't confuse its promise with precision.

You don't need AI to tell you that online brand visibility and discovery are shifting from search to the machine layer. This is the invisible space where models retrieve, rank and synthesise before they answer. That's where brand visability is determined.

GEO tracking is useful research. It shows patterns, signals and model drift, but it isn't ready for investment and decision-making for reputation management.

Treat any single "share of model" number with caution until three gaps close: clear definitions of what's counted, repeatability across runs, and the link to actual outcomes.

Until then, by all means run GEO for discovery. Show your method. Publish what you tested and what you found and be explicit about uncertainty. Signals and standards will follow if we're transparent.

The brief for communicators hasn't changed: invest in authority and trust. In practice, that means credible earned coverage, durable reference entries, and technically sound owned content that's easy to cite and hard to misinterpret.

That's best practice public relations.

Stephen Waddington, professional advisor and PhD researcher, Wadds Inc.

PR Agency One

Executive summary

GEO is here - but the evidence is chaotic

Generative Engine Optimisation (GEO) has arrived. It is already reshaping how information appears in search and conversation. Yet, as an emerging discipline, its evidence base is full of noise.

We are pro-GEO and pro-AI. At PR Agency One we believe this space deserves serious attention, investment and standards

We also recognise the bravery of agencies, tech vendors and analysts who have published early research, putting untested ideas into the public domain and advancing the debate.

But the results are wildly inconsistent. Across the literature, published figures on "what drives visibility in AI answers" differ by orders of magnitude.

- PR-led studies (Golin, Hard Numbers, Edelman, Muck Rack and yes even ourselves, PR Agency One) highlight the role of earned and journalistic sources.
- SEO-led studies (Ahrefs, SE Ranking, BrightEdge) emphasise owned and reference domains such as Wikipedia, YouTube and brand sites.

Both lenses reveal part of the truth, but neither provides a universal constant.

Early GEO measurement is valuable R&D, the sandbox where our industry learns what sticks, but calling it performance data risks misleading clients and misdirecting budgets.

In practice, that means treating early GEO tracking as research, not reporting

Experiment boldly - but label it clearly. Early GEO tracking is R&D, not ROI.

Why studies disagree

Three core issues explain the contradictions:

1. Different measurement frames. PR studies classify by media type and aggregate the long tail of coverage; SEO studies rank domains, which overemphasises a handful of global giants.

2. Different surfaces and modes. Google's AI Overviews, Perplexity and ChatGPT all behave differently, and their user interfaces don't always match their developer APIs.

3. Combinatorial overload. With billions of possible prompt variations, any "sample" of a few dozen queries is statistically meaningless.
(In short: different inputs, different rules, and too many possibilities for any single metric to hold steady.)

This report in numbers

Scattered throughout this report are references to public and proprietary data. This page summarises the key figures and sources at a glance.

30+
Independent GEO studies reviewed across PR, SEO and analytics disciplines between January and October 2024.

50 percentage-point swing
Range between published estimates of earned-media dominance in AI citations (Edelman 65%, Muck Rack 80%, Ahrefs 70%, SE Ranking 62%).

38 %
Average overlap between sources returned via ChatGPT's public UI and its API (Penn 2024; Otterly 2024) – the rest diverged.

4 quadrillion
Theoretical variations in a 12-word prompt where each word has 20 alternatives. Even one test per second would take >100 000 years.

0.0000001 %
Proportion of the English-language prompt universe represented by the largest public GEO sample sets (⩽5 000 prompts).

3 of 32
Number of studies that attempted to link AI visibility data to downstream brand outcomes; none produced statistically significant correlation.

55 – 70 %
Share of total AI citations attributed to earned-media outlets when the long tail is included, versus ⩽15 % from the top-ten mega-domains.

< 50 %
Of published reports disclosed their prompt lists, model versions or run counts – the main reason replication is almost impossible.

These figures are not an indictment of GEO measurement, they're a reminder that the field is still experimental. As new data and telemetry emerge, these baselines will evolve.

The problem with “share of model”

“Share of model” ie the idea of expressing brand visibility inside LLMs as a single percentage, is appealing in theory but unfit for purpose today.

It fails three basic tests:

- **Definition.** Most tools never explain whether they count mentions, citations or recommendations.
- **Reliability.** The same prompt can yield different answers seconds apart.
- **Outcome linkage.** No study has yet shown a causal link between AI visibility and business results.

Until those gaps are fixed, “share of model” should stay in the R&D lab - a useful experiment for pattern-spotting, not a KPI for reporting

.

The only reliable exception is where visibility is publicly verifiable, for example, Google’s AI Overviews, where inclusion and citations can be tracked over time.

Early visibility tracking has diagnostic value. Watching how a fixed set of prompts evolves over time can reveal model drift, shifting source preferences or technical artefacts.

Those patterns are worth observing, as long as we’re honest that they measure the system, not the audience.

why sampling prompts is futile

Suppose we test a 12-word prompt where each word has just 20 plausible alternatives.

That's 20^{12} = 4,096,000,000,000,000 variations - over four quadrillion possible prompts.

In human terms: even if you ran a new test every second, it would take more than 100,000 years to cover them all.

And that ignores follow-up questions, file uploads, regional settings and time-based model updates.

A single "share" number can never represent such an enormous, shifting universe.

A quick acid test for any "share of model" dashboard

Ask the vendor five simple questions:

1. **What is your denominator?**
   Show the full prompt universe you're sampling and its size.

2. **Is it the UI?**
   Prove that API results match what users actually see.

3. **Where is the variance?**
   Show repeat-run stability and confidence ranges.

4. **What are you counting?**
   Citations, mentions or recommendations - and for which intents?

5. **What moved in the real world?**
   Link results to brand or commercial outcomes.

If any answer is missing, you are not looking at a KPI.

What can be measured now

For now, GEO measurement should focus on what is observable:

1. **Visible AI surfaces** – such as AI Overviews and other answer panels users actually see.
2. **Source patterns** – classifying citations across earned, owned, shared and reference media, weighting the long tail of smaller outlets.
3. **UI validation** – checking that measurements match the consumer interface, not lab-only API data.
4. **Outcome linkage** – correlating these findings with real brand and commercial metrics.

Note: Alongside these observable surfaces, testing can serve as an R&D tool to track how models evolve, provided results are clearly labelled as experimental insight, not performance evidence.

Anything outside this evidence base is speculation.

A disciplined return to first principles

Until credible proxies and telemetry exist for AI visibility, the most effective approach is to double down on what is measurable:

- Classic brand and reputation tracking
- Media output analysis (coverage, tone, sentiment, message delivery, visibility)
- Attribution and outcomes through proven frameworks such as OneEval

These are not relics of a pre-AI age - they are reliable, reproducible metrics grounded in a known universe of data.

Real journalists, real publications, real audiences.

In this sense, GEO measurement is temporarily a return to the old school: evidence-led, transparent and benchmarked against reality.

Our position

GEO is real and worth engaging with - but it is too early for standardised KPIs.
Our recommendations:

- Measure only what users can see.
- Disclose methods and uncertainty.
- Anchor all GEO work in brand, reputation and commercial outcomes.
- Treat “share of model” as exploratory until new proxies and standards emerge.

When the technology and data catch up, measurement will evolve.
Until then, communicators should stay focused on the evidence that still matters most - what people actually see, read and remember.

What we recommend until the tech and standards catch up

A. Principles

- Reproducibility first. Publish prompt banks, model names, versions, interfaces and run counts. Repeat tests and report variation.
- Context over certainty. Show ranges and medians by model and query type; avoid single-point claims.
- Holistic inputs. Keep investing in earned authority, credible reference entries, high-quality owned content and sound technical SEO. All four levers matter.

B. A checklist for any AI visibility number

- Which surface was measured, and was it the UI?
- What models, versions and modes were used, and when?
- Exact prompt set - were context files, memory or browsing on?
- How many runs per prompt, and what variance was observed?
- How were sources classified - did you include the long tail of earned media?
- What outcome metrics changed alongside?

C. Measurement tools

Use tools, like our OneEval Brand, Reputation and Commercial, to anchor GEO activity to outcomes, with an AI surfaces module tracking inclusion and source patterns across assistants.

Where we land

We are enthusiastic about GEO and experimentation, but cautious about claims of precision..

A single black-box number cannot summarise an infinite space.

The pragmatic path is simple:

- measure only what users can see,
- build authority in sources models already trust,
- disclose methods and uncertainty,
- and link every observation to real outcomes

As standards mature and platforms expose better telemetry, we will revisit whether “share of model” can graduate from an idea to a reliable metric.

What the interfaces really show — and why API-only studies mislead

The UI and API don't tell the same story

Anyone who has compared ChatGPT's web interface with its API results knows they differ.

The difference sounds minor, but it explains why so many "AI visibility" experiments contradict one another.

When people use chat assistants in a browser, the system adds invisible layers - memory of past conversations, tone adjustments, safety filters, even editorial style.

It may also trigger tools such as **browsing** (fetching live web pages), **code execution**, **or deep research,** each of which expands how the model finds information.

The API - the interface developers use to send text to the model programmatically - strips most of that away. It is faster and cheaper, but not what the public sees.

In practice, API results show how the model behaves in the lab; UI results show what real people experience. They are not interchangeable.

Other factors widen the gap:

- **Memory and context.** Real users have ongoing chats; APIs start from a blank slate every time.
- **Version drift.** Public interfaces quietly update to new model versions; API testers may still be querying the old one.
- **Randomness.** LLMs introduce controlled randomness (called temperature). Even a tiny change can alter an answer. Few studies re-run prompts to measure that variation.

Bottom line, API-only data describes controlled conditions, not consumer reality. Unless a study compares its API results with what users actually see, it is not describing real-world behaviour.

Different assistants, different instincts

Large language models (LLMs) are not identical twins. Each has its own training data, retrieval system and editorial bias - which means they cite and phrase information differently.

- **ChatGPT** favours encyclopaedic and mainstream news sources. When its browsing tool is switched on, it may also quote directly from brand websites for highly specific, navigational queries.
- **Google's AI Overviews** draw mainly on web pages that already rank well in Google Search, but they also weave in forums and YouTube clips for colour. In other words, traditional SEO signals still matter here.
- **Perplexity** mixes formal references with community discussions, Reddit threads and video content.
- **Bing Copilot** and **Claude** vary by mode. Some versions display citations; others summarise without attribution, which hides the underlying sources entirely.

And in plain English, each platform has its own personality and source preferences.

Comparing them without noting those differences is like comparing TV ratings with podcast downloads, both measure attention, but in incompatible ways.

Why “top-domain” tables miss the point

Many SEO-oriented studies list the ten most-cited domains in AI answers.

Those tables always look impressive - Wikipedia, YouTube, Reddit and other mega-sites dominate the charts - but they hide the real story.

These league tables show only the head of the distribution: the same few giants that appear everywhere.

They ignore the long tail of thousands of smaller publications, trade journals and blogs that collectively provide far more citations.

PR-side studies, which group results by media type (earned, owned, shared, reference), make that tail visible.

When you count every small outlet once, earned media clearly outweighs the top-ten super-domains.

In short, SEO leaderboards aren’t wrong, just incomplete.

They describe the big head of the curve; PR-style analysis captures the long tail.

You need both to understand how AI assistants actually source their information.

Why “share of model” isn’t ready for prime time… yet.

The phrase share of model has become fashionable - the idea that we can express a brand’s visibility inside LLMs as a single percentage.

It sounds elegant, but the maths and data infrastructure simply aren’t there yet.

To qualify as a decision-grade metric, any number needs three things:

1. A clear definition. Are we counting citations, text mentions, or recommendations? Most studies never say.

2. Reliability. If you ask the same prompt twice and get different answers, the measurement isn’t stable.

3. Outcome linkage. A useful KPI should connect to real-world effects such as awareness, traffic or sales. None of the published “share of model” figures do.

Until those problems are solved, treat it as a thought experiment, not a KPI.

The only partial exception is where the AI output is visible and verifiable - such as Google’s AI Overviews - where inclusion and citation counts can be checked like any other search feature.

The missing denominator. Prompts aren’t keywords

Search engines have keywords; AI chat has prompts.

In classic search, keyword data tells us what people ask and how often - the denominator that makes “market share” calculations possible.

In AI chat, that denominator doesn’t exist. Prompts are private, multi-sentence and context-dependent.

Let’s unpack that:

- Prompts are elastic. Real users don’t type two words; they write paragraphs. A single intent can be phrased thousands of ways.
- Context stacks. Conversations carry memory, uploaded files and previous questions that silently change each new answer.
- Privacy by design. There is no public “prompt volume” database like Google’s keyword planner.
- Variable models. Switching browsing, safety filters or regional settings changes outputs.
- Interface differences. API tests rarely mirror what the consumer UI displays.

Bottom line, without a denominator, any “share” number is guesswork.

You can measure visible AI surfaces for defined prompts, but you cannot claim a global market share of an invisible, shifting universe.

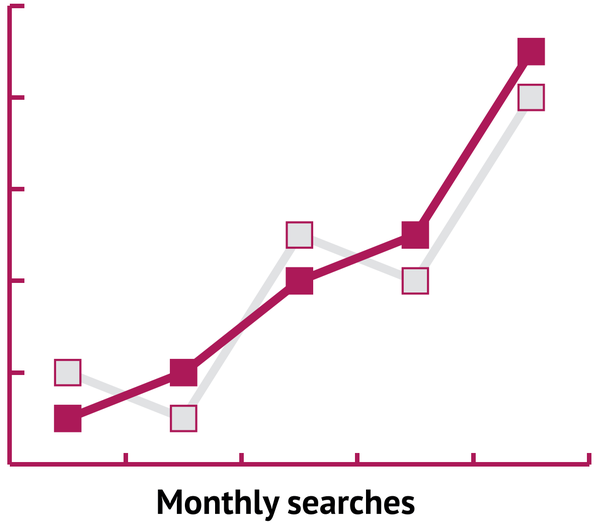
> **Without prompt volumes there is no denominator.**
> **A ‘share’ of something you cannot count is not a metric - it is a guess.**

PR AgencyOne

Keywords vs Prompts – Why ‘share’ collapses

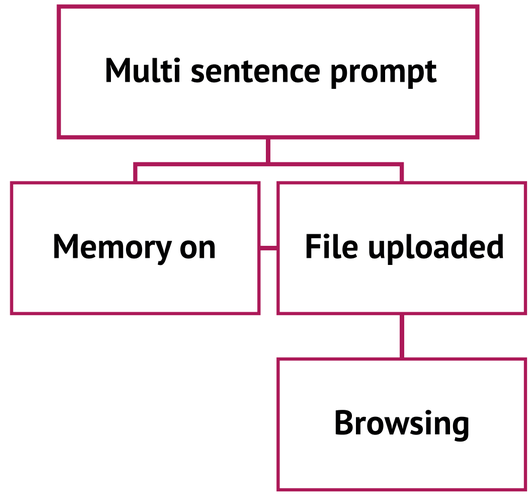
Keyword world

Short queries, public volumes, fixed results page.



Prompt world

Multi-sentence asks, private context, memory, files, browsing.



Search gives volumes. Prompts don’t. No volumes = no denominator = no “share”

PR AgencyOne

Who is saying what

The GEO conversation currently splits into two camps:

- **PR and communications agencies** such as Golin and Edelman emphasise that AI assistants overwhelmingly cite earned and journalistic sources. Muck Rack's dataset shows that most links surfaced by AI are unpaid, editorial content. (This is something our own research has shown and as a PR person I am inclined to agree with)

- **SEO and analytics vendors** such as Ahrefs and SE Ranking concentrate on domain-level citation patterns, often within Google's AI search. Profound's cross-platform analysis maps how those preferences vary by assistant.

- **Independent analysts** including Christopher S. Penn and tools like Otterly highlight how sensitive results are to prompt wording - and how wide the gap is between API and UI outputs.

Together, these perspectives paint a fuller picture: GEO research is vibrant, but fragmented.

Detailed findings and what we actually did

How we built the evidence base
To understand why so many "AI visibility" studies disagree, we reviewed every publicly available piece of research published since January 2024 that tried to quantify how AI assistants surface brands or information.

We included studies from PR agencies and vendors (Edelman, Golin, Hard Numbers, Muck Rack, Profound and yes our own PR Agency One study) and from SEO platforms (Ahrefs, SE Ranking, BrightEdge and others).

Each was evaluated on three criteria:

1. **Transparency –** Did the authors explain how they gathered their results, which models or versions were tested, and whether they used the public interface (UI) or a developer interface (API)?

2. **Reproducibility –** Could another analyst repeat the same tests and get similar answers?

3. **Scope –** How many assistants, prompts and industries were covered?

In total, we reviewed more than 30 substantial reports analysing ChatGPT, Google's AI Overviews, Perplexity, Bing Copilot and Claude.

Most studies were observational: they asked questions of the AI interfaces and counted which sources were cited.

However, very few published their prompt lists or re-ran the same prompts to check consistency. That lack of methodological disclosure explains much of the variation in published results.

(Plain English: most teams tested, but few explained exactly how - so we can't know whether their data would repeat the same way twice.)

What we learned about the UI vs API gap

A recurring theme is that **API testing doesn't match human experience.**
When researchers query AI models via the API, they often bypass the invisible layers that influence what people see - such as memory of past chats, browsing, regional filters or safety settings.

These experiments show how models behave in controlled conditions - not how people discover information. It's the difference between studying the weather in a wind tunnel and the climate outside. Useful for pattern-spotting, dangerous if mistaken for consumer insight.

Tests run in the public interfaces, the way real users interact, often produce different citations or even different brand mentions compared with API tests using the same words.

In practice, this means any "share-of-model" score built solely from API calls describes **laboratory conditions**, not the real world.

Until interfaces and APIs are aligned or standardised, **UI-based observations** are the only meaningful benchmark for brand visibility.

(Think of it like testing a car engine on a stand versus driving it on the road - both generate data, but only one reflects real performance.)

What can safely be concluded

From all the reviewed material, five points are clear:

1. **UI and API results are not interchangeable.**
   Treat API-only data as experimental until it has been validated against what users actually see in the interface.

2. **Platform behaviour varies.**
   ChatGPT favours reference and news sources; Google's AI combines search and social signals; Perplexity gives more weight to community content.

3. **Results are volatile.**
   Even identical prompts can yield different answers depending on timing, model version, or small randomisation factors built into the AI.

4. **Leaderboard bias is real.**
   A few mega-domains dominate headline charts, but the combined influence of smaller earned outlets is far greater in total.

5. **Outcome linkage is rare.**
   Few studies connect AI-side numbers to tangible business results. Until that happens, treat those figures as signals, not KPIs.

The long-tail adjustment, in plain English

To correct for the skew caused by mega-domains, we calculated a tail-weighted earned share: combining the small proportion of citations from top global sites with the much larger volume from smaller outlets.

When weighted this way, earned media typically represents the majority of all AI citations - even though no single outlet appears in the top ten.

What this means for communicators:
Being cited once in the right publication still matters, and doing it consistently across many smaller outlets matters even more.

Visibility in AI is not about one headline domain; it is about cumulative authority built through repeated, credible mentions.

Back to what we can measure

- For now, the most effective way to track brand impact in the AI era is to return to proven disciplines:
- Reputation tracking and sentiment analysis
- Media output and message delivery audits
- Google Analytics data
- Outcome linkage through our tools like OneEval Brand, Reputation and Commercial
- These methods may seem “old school”, but they remain the only frameworks grounded in transparent, verifiable data.
- GEO will eventually have its own standards, but until then, disciplined use of existing evaluation tools offers the most reliable picture of how communication drives visibility and trust.

Conclusion

GEO is real and important, and early testing is essential to build the evidence base, but we should be clear about what's signal and what's noise.

Until the data infrastructure matures, the most responsible course is to blend careful experimentation with the proven rigour of brand and reputation tracking

Current visibility-tracking tools lack reliable denominators, consistent interfaces and outcome linkage.

Rather than chasing phantom metrics, the pragmatic route is to:

1. Measure what users actually see - observable AI surfaces such as Google's AI Overviews.

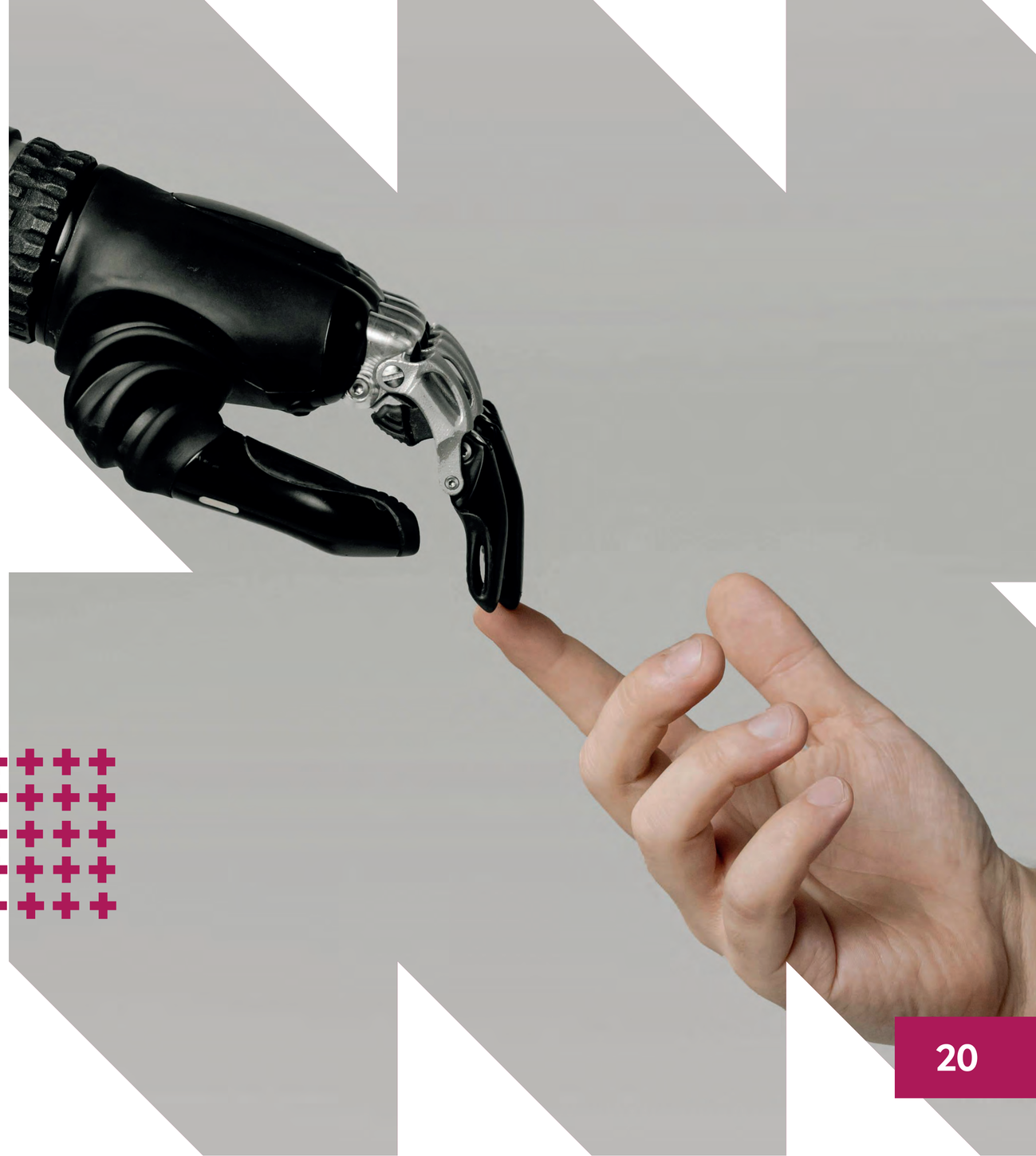
2. Disclose methods and variance - so results are credible and reproducible.

3. Anchor everything in proven frameworks - brand and reputation tracking, sentiment analysis, message delivery, and business outcomes through systems such as OneEval Brand, OneEval Reputation, and OneEval Commercial.

In time, new proxies will emerge to make AI visibility genuinely measurable.

Until then, communicators are best served by evidence, not speculation, ie building authority through credible coverage, and proving impact through the data that still matters most: real-world reputation and results.

The more we experiment transparently, and admit what we don't yet know, the faster credible GEO standards will emerge.



PR AgencyOne

About the author

James Crawford FPRCA

Managing Director, PR Agency One

James Crawford is the founder and Managing Director of **PR Agency One**, an award-winning communications consultancy based in Manchester and London.
He has more than 20 years' experience in reputation management, corporate communications and measurement, working with international brands across technology, finance, healthcare, built environment and consumer sectors.

Under his leadership, PR Agency One has won multiple AMEC, PRCA and CIPR awards, including seven Agency of the Year titles.

The agency's proprietary **OneEval** measurement suite – combining *OneEval Brand, OneEval Reputation* and *OneEval Commercial* – has become one of the UK's most recognised models for linking PR performance to commercial outcomes.
James is a Board Director at **AMEC**, the global association for communications measurement and evaluation, where he helps shape standards for data-driven PR.
He writes regularly on Creative Effectiveness, analytics and the evolving relationship between earned media, AI and search.

When not writing or presenting, he divides his time between Manchester and London, balancing family life, techno vinyl collecting and a stubborn belief that PR should always be measurable.

Who we are

PR Agency One is an independent consultancy that blends creativity, rigour and measurability under its central proposition of **Creative Effectiveness.**

Its in-house evaluation platform **OneEval** enables clients to track the brand, reputation and commercial impact of communications activity against business goals.

With specialist divisions spanning Consumer & Retail, Healthcare, Finance, Technology and the Built Environment, the agency operates internationally through One Network, a 26-country partner alliance.

Find out more at www.pragencyone.co.uk

Acknowledgements

Thanks to colleagues at PR Agency One and contributors across AMEC, the PRCA, and the wider GEO and SEO research community who have tested, challenged and refined the ideas presented in this paper.

Particular thanks to those agencies and analysts who have shared prompt datasets and UI testing results in public, genuine progress only happens when evidence is open.

Our award-winning work

Over 80 industry awards including:

Tech agency of the year • B2B agency of the year • Awards in search, PR, marketing and evaluation

PRmoment Awards 2019 Independent Agency of the Year

PRmoment Awards 2019 Technology PR Agency of the Year

CIPR PRIDE AWARDS 2019 GOLD WINNER

amec AWARDS | 2019 WINNER

CIPR excellence AWARDS 2019 WINNER

PRCA DIGITAL AWARDS 2019 WINNER

PR AgencyOne

Next steps...

Ready to get started?
Speak to the PR experts today.

enquiries@pragencyone.co.uk

London 020 3092 1446

Manchester 0161 871 9140

PR AgencyOne