Machines have an opinion of your brand.

Here's how to read it — and shape it.

Robert Maynard
Founder
AnswerShare

Acknowledgment

Portions of this paper were developed with the assistance of generative AI tools used for drafting, editing, synthesis, and analytical exploration.

λNPS™ — a measure of AI sentiment towards a brand

ANSWERSHARE RESEARCH · WHITEPAPER

PRE-PUBLICATION DRAFT FOR PEER REVIEW

Abstract

By the time a user asks an AI assistant about a company, the machine has already formed a position. That position is assembled — quietly, at scale — from the public corpus written *about* the brand: reviews on Trustpilot and Amazon, complaint threads on the Better Business Bureau, forum posts on Reddit, transcripts of YouTube testimonials.

The machine is not ranking the brand against competitors. It is averaging an opinion, then passing that opinion to the next ten thousand users who ask.

This paper proposes **λNPS (lambda-NPS)**, a measurement system that asks five canonical NPS questions of four AI engines directly, returning what each machine reveals under controlled prompting. λNPS is the live signal — the reputation the machines are presently expressing about a brand to the next ten thousand users who ask.

To make λNPS interpretable, the paper pairs it with **μNPS™ (Modeled Net Promoter Score)**, a companion measurement that reconstructs the classical -100..+100 NPS distribution developed by Bain et. al, applied to machine opinion based on the open-market corpus extant the brand.

The gap between the two is denoted **ΔNPS™** and is the framework's central diagnostic: it tells a brand whether the corpus underneath is being faithfully translated, or whether the retrieval and weighting layer is distorting the signal in transit.

The framework is not presented as a universal truth metric, nor as a direct measurement of latent model state. It is proposed as an operational measurement system intended to:

- measure machine-expressed reputation under controlled prompting (λNPS),
- quantify the corpus substrate beneath it using weighted public signals (μNPS),
- and isolate the translation-layer gap between the two (ΔNPS) as a versioned, reproducible diagnostic.

λNPS measures the brand the AI is currently describing. ΔNPS measures the distance between that description and the corpus underneath it — a measurable artifact, and one we propose as a leading indicator of where AI-mediated brand reputation appears to be moving. This document is a pre-publication draft intended for technical, statistical, and methodological review.

1. Introduction

> *AI doesn't rank you. It forms an opinion of you. Then it tells the next 10,000 users what it thinks.*

This is not a metaphor, and it is not a marketing line. It is the literal mechanic.

When a user types "is {BRAND} trustworthy?" into ChatGPT, Perplexity, Gemini, or Claude, the model is not consulting an index, sorting candidates by relevance, and presenting a ranked list of links. It is generating language — a continuation, conditioned on everything the model has read about the brand during training, and on whatever live search results it chose to retrieve at inference time.

The output reads like a measured judgment because the model has been trained to produce measured judgments. Underneath that polish is something closer to averaged sentiment, weighted by what the model treats as credible.

Britney Muller, who has spent the better part of two decades thinking about how search systems weight signals — first at Moz running Whiteboard Friday, then at Hugging Face on the AI side — puts it plainly. *"There is no algorithm,"* she told the AI SEO Show in late 2025 (Muller, *AI SEO Show*, 1 November 2025, https://www.youtube.com/watch?v=l4fIHPtjIMY). *"There are weights. There are no rankings. These are probabilities. It's a nondeterministic system."*

Marketers who keep using the vocabulary of SEO when talking about LLMs are making a category error. There is no first-place finisher in an AI answer. There is a probability distribution, sampled at temperature, that produces one continuation among many possible ones — and most of the time, the continuations rhyme.

That rhyming is the opinion. Run the same prompt twice and the wording shifts; the sentiment usually does not. Run it across four different engines and you start to see the contour of what the public corpus

has taught them. The shape doesn't change much from week to week. That is the thing worth measuring.

Two independent studies in the last twelve months have made the same point with empirics rather than vocabulary. Surfer's 2026 trends report, drawing on 289,000+ URLs, found that the correlation between third-party brand mentions and AI recommendations sits at roughly **0.41 Spearman** (Kohli, *SEO Trends 2026*, Surfer SEO, 8 May 2026, https://surferseo.com/blog/seo-trends-2026/) — meaningfully higher than the on-page tactics Surfer measured in the same dataset, where early keyword placement showed near-zero correlation with citation.

In plain English: how often a brand is mentioned across the open web, in natural language, by sources the machine considers authoritative, appears to predict whether the machine will recommend it. The own-website surface matters less than it used to. The corpus *about* the brand matters more. Correlation is not causation, and the 0.41 figure is one vendor's empirical read on one dataset — we hold it as one vendor's read, not as proof of a universal law.

Kevin Indig's Growth Memo study, published in April 2026, examined 3,981 domains across 115 prompts in 14 countries — and found that 61.7% of citations are ghost citations (Indig, The Ghost Citation Problem, Growth Memo, 20 April 2026, https://www.growth-memo.com/p/the-ghost-citation-problem): the AI drew on a domain's content without naming the brand. Three out of every five times a brand's content shaped an AI answer, its name did not appear in that answer. The machine is forming opinions using brand material without telling the user where it came from. That is the layer beneath the visible answer, and it appears to be where reputation now gets decided.

This is not "search with a chatbot interface." This is a different sort of system, and the vocabulary that worked for search — rankings, positions, SERPs — does not describe it. The closer analogy is reputation.

You walk into a room and the people there have already heard things about you, from sources you do not always know, weighted by trust you did not always earn. They form an opinion before you open your mouth. The machine works the same way. The room is much larger.

This paper proposes that AI-mediated reputation is measurable, that corpus sentiment and machine-generated sentiment can meaningfully diverge, and that the delta between them is itself diagnostically useful.

1.1 Notation

This paper uses a three-symbol notation throughout. It is introduced here so that the rest of the document reads cleanly:

μNPS = modeled reputation reconstructed from public corpus signals
λNPS = machine-expressed reputation measured from AI-generated outputs
$\Delta NPS = \lambda NPS - \mu NPS$

The notation distinguishes modeled corpus reputation from machine-expressed output and isolates the diagnostic gap between them. The reader meets μNPS first as the corpus-side measurement (§3), λNPS in the middle as the machine-side companion (§5), and ΔNPS as the analytical payoff that emerges from running both side by side.

The architecture as a diagram:

Public Corpus
reviews · forums · BBB · Reddit · video
↓
μNPS
modeled corpus reputation
↓
AI Retrieval / Weighting Layer
↓
λNPS
machine-expressed reputation
↓
ΔNPS = λNPS − μNPS
translation-layer divergence

λNPS is the live signal — what the machine is presently saying about a brand. μNPS is the corpus substrate beneath it. The retrieval/weighting layer is the transform between them. ΔNPS is what falls out when the two are compared, and is where most of the diagnostic work happens. The rest of the paper unfolds each layer in turn.

2. The Corpus Problem

> *You can rewrite your homepage tonight. You cannot rewrite the 18,000 reviews already written about you. AI reads both — and weights them differently.*

The corpus that shapes a machine's opinion of a brand is not solely the brand's website. It is the public material *about* the brand, published by other people, on platforms the brand does not control. Trustpilot. Amazon. The Better Business Bureau. Reddit. YouTube. Industry forums. Comparison sites. Local subreddits where customers vent at midnight. The transcripts of podcast interviews where the founder did or did not handle a hostile question well.

All of it readable, all of it aggregable, all of it weighted by signals the machine considers — recency, source authority, sentiment, specificity, intensity.

The reason this matters is structural. Most of these platforms publish in formats designed for machine consumption. Trustpilot, for instance, exposes its reviews in JSON-LD `Review` schema — title, body,

rating, author, date — the same structured-data format Google has spent fifteen years asking publishers to adopt.

The machine doesn't need to scrape and infer. The machine reads the structured field, the structured field has a star rating, and the star rating walks straight into the model's weighted average. Amazon does the same. The BBB does the same. The web is structured around the idea that reviews are first-class data, and the AI systems appear to benefit from that structure more than any search engine ever did.

This is the control inversion that makes brand work harder than it used to be. A brand owns its domain. A brand does not own the corpus. A flawless product page published tonight does not erase a Trustpilot review from a customer who waited fourteen minutes on hold and never recovered emotionally.

The machine sees both. The machine appears to weigh them differently. And the part the brand didn't write often weighs more, because it is socially expensive to produce — a stranger took the time to write 400 words about the brand for no compensation — and the machine treats that cost as a credibility signal.

Muller, in two separate venues — her AI SEO Show appearance and the Elumynt UpArrow podcast (Muller, *Elumynt UpArrow podcast*, 2025, https://www.elumynt.com/podcast/actionable-ai-for-marketers-the-human-in-the-loop-with-britney-muller) — has used the same shorthand: *"Brand mentions are the new backlinks."* The pretrained layer of an LLM is, in her framing, "your mediocre generating engine," the one that produces the statistical average of everything it has read about a thing. The more times a brand appears in that corpus, in relevant context, the higher the probability that the model surfaces it when next asked.

This is not opinion. It is mechanics, consistent with Surfer's 0.41 Spearman number on the recommendation side (Kohli, Surfer SEO, 2026) and with Indig's 61.7% ghost-citation rate on the citation side (Indig, Growth Memo, 2026). The corpus shapes the output. The corpus is where the work is.

Aleyda Solís, on *Humans of Martech* in January 2026, made the related point about treatment. *"Audit your share of voice and sentiment within each topic versus competitors,"* she argued, *"and dig into whether you are mentioned in a negative light."* (Solís, *Humans of Martech*, episode 202, 13 January 2026, https://humansofmartech.com/2026/01/13/202-aleyda-solis-ai-search-crawlability/) AI search, in her framing, "must be treated as a branding channel, not only a performance channel. Inclusion, share of voice, and sentiment in answers matter even when they generate no direct click or referral traffic."

The thing being measured is not traffic; it is presence-in-an-opinion. Solís separately publishes a checklist that includes, as one of its ten steps: *"Monitor your brand mentions, sentiment, and links separately for each major AI search platform."* (Solís, *The 10 Steps AI Search Content Optimization Checklist*, aleydasolis.com, 25 July 2025, https://www.aleydasolis.com/en/ai-search/ai-search-optimization-checklist/) That is, at a higher resolution, the work this paper proposes to formalize.

The remaining question is how. If the corpus is the input and the machine's opinion is the output, what is the measurement layer between them? Classical NPS — the survey — was designed to measure a

different system. It asks the customer on a 0-to-10 scale and produces a -100 to +100 score. It is honest, it is well-studied, and it is unfit for the new problem because it only measures what the customer says when asked. The new measurement has to read what the customer already said, to the machine, in public.

3. Proposal: µNPS

NPS is what your customers tell you when you ask. µNPS is what they already told the internet.

The Modeled Net Promoter Score developed by Bain et al, reconstructs the classical NPS distribution from open-market signals. It uses the same -100 to +100 range, the same promoter/passive/detractor buckets, the same arithmetic skeleton. What changes is the input source. Instead of soliciting responses, µNPS reads them where the customer already left them.

The formula is preserved in its standard shape:

```
µNPS = (promoter_weighted_pct − detractor_weighted_pct) × 100
```

What changes is the weighting. Not every signal in the corpus carries equal credibility, and pretending it does would produce a number that does not survive contact with how the AI engines themselves appear to weight sources. Methodology version 1 of µNPS uses five signal types, each with a tunable weight:

SIGNAL TYPE	WEIGHT	RATIONALE
verified_review	1.0	Verified platform review (Trustpilot, Amazon, BBB) — the baseline unit
long_form_review	1.5	Reviews exceeding 200 characters — more signal-dense, more retrieval-likely
video_testimonial	2.0	Highest-credibility signal — costly to produce, hard to fake
forum_complaint	1.2	Reddit / BBB complaint threads — public, attached to identity
social_mention	0.5	Drive-by social references — lowest confidence per unit

These weights are heuristic priors, not empirically finalized constants. They are informed by observed retrieval behavior, public AI-search studies, operational testing, and corpus characteristics. Future methodology versions may materially change them.

The weights themselves are an argument, and one worth pausing on. A video testimonial is weighted four times a social mention because it took multiple orders of magnitude more effort to produce. A 250-word Trustpilot review is weighted one-and-a-half times a short one because the model can extract

more from it — sentiment, specificity, intensity, context — and because the AI engines themselves appear to draw on longer, more substantive content at higher rates.

Surfer's research, again drawing on the 289K-URL dataset (Kohli, Surfer SEO, 2026), found that pages with ten or more discrete facts get cited at roughly twice the rate of fact-thin pages. Long-form content appears to compound in the same way at the corpus level: more substance per unit, more retrieval per unit, more weight per unit.

That weighting choice foreshapes the strategy in §7. If a brand is trying to move its score, the brand should be doing the things that produce high-weight signals.

One more thing about the schema: every score row records its `methodology_version`. When v2 weights are introduced — when, say, we learn that video testimonials should be weighted 2.5 rather than 2.0, or that BBB complaints should be split into pre- and post-response classes — those weights apply only to scores computed under v2. Historical v1 scores are not retroactively rewritten.

This is the discipline that distinguishes a measurement system from a moving target. If a brand watches its µNPS climb from +12 to +47 over a year, the brand should be able to verify that climb was real movement in the corpus and not a quiet weight adjustment behind the scenes.

The µNPS score is the corpus side of the equation. It is, by construction, a measurement of *input* — what the public is publishing about a brand, weighted by credibility, normalized to a familiar range. The next question is what the machine actually *does* with that input.

4. Sentiment Classification

µNPS uses LLM-assisted sentiment classification rather than rule-based lexicons alone. The rationale is contextual interpretation. Consider the following review:

> *"The product failed initially, but support resolved the issue quickly and I would buy from them again."*

Rule-based systems may overweight negative keywords ("failed") while underweighting recovery and advocacy language ("resolved," "would buy from them again"). VADER, the most-cited rule-based sentiment library, scores this sentence negative because it weighs "failed" without weighing the recovery clause that follows.

An LLM with reasonable temperature reads the sentence the way a person would: a crash, a fast recovery, and a customer who is now a promoter. The frame is "this is the kind of company I would buy from again." The system pays the per-call cost for that reading because the alternative — calling promoters detractors — is unrecoverable downstream.

The proposed v1 system therefore uses:

- continuous sentiment scoring from -1.0 to +1.0,
- promoter/passive/detractor bucketing at ±0.15,
- and contextual classification using Claude Haiku 4.5 with prompt caching on the system rubric — the cached system prompt is billed at a fraction of normal input tokens on a 5-minute TTL, which is the cost-control mechanism that makes per-signal LLM classification economically viable at corpus scale. Periodic audits compare Haiku outputs against the current Anthropic frontier (Claude Opus 4.7 and Claude Sonnet 4.6) to catch classifier drift.

The use of Haiku should not be interpreted as an assumption that smaller models are inherently sufficient or universally accurate for sentiment classification. The methodology instead treats the classifier as a versioned measurement instrument whose behavior must itself be audited and calibrated over time. Current implementation pairs production-scale Haiku classification with periodic frontier-model comparison audits and planned human-labeled benchmark sets.

Current thresholds:

- **promoter:** sentiment_score > +0.15
- **passive:** -0.15 ≤ sentiment_score ≤ +0.15
- **detractor:** sentiment_score < -0.15

The boundary at ±0.15 is the operational definition of the "passive" band. A score of +0.10 is passive; a score of +0.20 is a promoter. This threshold matters because it controls how aggressive the score is — narrowing the passive band increases volatility, widening it dampens movement. Version 1 uses ±0.15 because empirically that appears to capture the natural break between "It worked, arrived on time" (functional, no advocacy) and "It worked well and I'd come back" (advocacy, even if mild). These thresholds are operational assumptions and may evolve.

When the Anthropic API is unavailable (rate limits, regional outage, expired key), the classifier falls back to a deterministic keyword-only heuristic. The heuristic counts hits against 27 promoter keywords and 33 detractor keywords, computes a normalized score from the differential, and caps the absolute score at ±0.9 — never reaching the full ±1.0 range that the LLM is allowed.

The cap is a confidence signal: a score that hits ±0.9 exactly is recognizably keyword-only and can be flagged downstream as lower-confidence than a full LLM read. The fallback is intentionally conservative; it produces *some* score for every signal so the system degrades gracefully, but it does not pretend to match LLM-quality reading.

Every signal also captures three deterministic modifiers from the raw text, independent of the LLM call: `caps_ratio` (uppercase-alpha character ratio, as a proxy for emotional intensity), `length_chars` (a substance proxy), and `has_specifics` (boolean — does the text contain a four-digit year, a date pattern, or a quoted phrase). These do not modify the score directly in v1, but they are persisted on every signal so future methodology versions can incorporate them — e.g., upweighting long, specific reviews relative to short emotional ones — without re-running the LLM.

5. Trust Priors

As large language models evolve from static, memory-driven systems toward live retrieval and grounded answer generation, trust increasingly appears to function as a foundational ranking and citation mechanism. This emerging concept may be understood as a "trust prior" — a probabilistic pre-existing confidence that an AI system assigns to a domain, platform, entity, or publisher before evaluating the specific content itself.

Human beings naturally use trust priors. Information published by the Mayo Clinic, the SEC, the Wall Street Journal, or Amazon begins with an inherent credibility advantage over an anonymous blog or unmoderated forum. AI systems appear to be developing similar behavior patterns. In an environment where retrieval systems must evaluate sources under latency constraints, hallucination risk, and grounding requirements, systems cannot practically treat all sources equally. They require heuristics that help estimate reliability rapidly and efficiently.

Trust priors likely emerge from the accumulation of many observable signals over time, including:

- semantic consistency
- structural clarity
- historical accuracy
- source grounding
- moderation quality
- retrieval efficiency
- cross-source corroboration
- entity resolution confidence
- contradiction frequency
- platform integrity

As AI systems increasingly rely on live retrieval and citation-aware generation, these priors may materially influence whether a source is selected, reranked, grounded, or ultimately incorporated into generated answers.

Importantly, trust priors are likely not binary. They are probably probabilistic and continuously adjusted through ongoing retrieval interactions. A domain that consistently produces grounded, semantically coherent, low-ambiguity information may strengthen its trust prior over time. Conversely, domains associated with inconsistency, manipulation, spam, or adversarial optimization may experience trust degradation.

This represents a fundamental shift from traditional search-era visibility models. Search engines historically optimized heavily for popularity signals such as backlinks, click behavior, and keyword

relevance. AI systems operating under grounded generation constraints appear increasingly incentivized to optimize for trust preservation, confidence, and ambiguity reduction.

In this environment, visibility may become less dependent on who is most discoverable and more dependent on who is most trusted.

6. λNPS

Trust priors are expressed operationally through λNPS: a probabilistic measure of an AI system's inferred willingness to retrieve, ground, cite, or recommend a source during answer generation. In this paper, we refer to this machine trust preference as λNPS.

Frontier AI systems increasingly rely on grounding, retrieval quality, source attribution, and trust-aware generation techniques to reduce hallucinations and improve reliability. Google, Anthropic, and recent RAG research all emphasize grounded retrieval, trustworthy attribution, and confidence-aware generation as foundational to modern AI system behavior (Google Cloud, How Vertex AI Grounding helps build more reliable models, Google Cloud Blog, https://cloud.google.com/blog/products/ai-machine-learning/how-vertex-ai-grounding-helps-build-more-reliable-models; Anthropic, Introducing Contextual Retrieval, anthropic.com/engineering, https://www.anthropic.com/engineering/contextual-retrieval; Song et al., Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse, arXiv:2409.11242, 17 September 2024, https://arxiv.org/abs/2409.11242).

AI systems will not recommend what they cannot confidently trust. λNPS is the measurable expression of that trust.

Trust is the concept. λNPS is the metric.

If µNPS reads +60 and λNPS reads +12, the resulting ΔNPS of −48 is the AI reading you but not believing you. A negative ΔNPS means the machine is judging the brand more harshly than the human corpus does — the translation layer is leaving signal on the table. That gap is what is worth measuring.

µNPS measures corpus input. λNPS attempts to measure machine output. The distinction is central to the framework.

> ***A NOTE ON THE λ SYMBOL.*** *The λ symbol is intended to denote machine-transformed output behavior rather than latent model state. λNPS measures generated responses under controlled prompting conditions, not internal model beliefs or hidden parameters.*

That caveat is load-bearing and worth restating. The internal state of a large language model is not directly observable, and this paper makes no claim to observe it. What λNPS measures is what the systems currently generate under controlled prompting conditions — the same questions, classified by the same rubric, run on a regular cadence. λNPS measures machine-expressed reputation output under controlled prompting conditions.

The current implementation runs five canonical NPS-aligned prompts against four AI systems:

1. *What's your opinion of {brand}?*
2. *Would you recommend {brand} to someone looking for their primary product or service?*
3. *What are {brand}'s strengths and weaknesses?*
4. *Is {brand} a trustworthy company? Explain.*
5. *What's the general consensus about {brand} online?*

Each prompt is run against four engines — Perplexity (Sonar, the production default backed by GPT-5.5 / Claude Sonnet 4.6 / Gemini 3.1 Pro in its auto-routing pool), OpenAI (GPT-5.5, the current ChatGPT default as of May 2026), Gemini (3.1 Pro, the current consumer frontier), and Claude (Opus 4.7, with prompt caching on the system prompt for cost control) — producing twenty responses per brand audit.

Each response is then classified independently by Claude Haiku into promoter / passive / detractor, with reasoning and confidence captured. The audit takes a few minutes and costs roughly $0.053 per brand: $0.025 for Perplexity calls, $0.005 for OpenAI, $0.0025 for Gemini, $0.001 for Claude generation, and $0.020 for the twenty classifications that follow. At that cost point, λNPS is not a quarterly study; it is something a brand can run weekly without thinking about it, which is the only cadence that makes a non-deterministic system trackable.

Different frontier models may produce materially different λNPS outputs for the same brand under identical prompts. The per-engine breakdown is the part that matters most, and it is a deliberate design choice. The aggregate λNPS — the median across all four engines — is a clean headline number, but the engines diverge.

Indig's *Ghost Citation Problem* (Indig, Growth Memo, 20 April 2026) found that Gemini names brands in 83.7% of appearances but cites them only 21.4% of the time; ChatGPT inverts the pattern, citing 87.0% but naming 20.7%. The dataset behind those numbers showed roughly 22% disagreement between engines on whether a brand was mentioned at all for the same query. That is not noise. That appears to be the engines reading the same corpus and arriving at materially different positions, because each is weighting the corpus differently.

A brand's λNPS on Perplexity can be +30 while its Gemini number is -15, and the difference may contain real information about which engine has weighted which signals heavily. This inter-engine disagreement — what we informally call **λ-divergence** — is itself a signal worth tracking. Solís's ten-step checklist (Solís, aleydasolis.com, 25 July 2025) specifies the same discipline directly: *"Monitor your brand mentions, sentiment, and links separately for each major AI search platform."* The four-engine breakdown is the operational form of that recommendation.

That diversity is what makes ΔNPS interesting. The two scores answer different questions:

- **μNPS** measures the input. What is the public saying about the brand?
- **λNPS** measures the output. What are the machines saying about the brand?

A brand with μNPS of +60 (the public, on net, likes them) and λNPS of +12 (the machines, on net, are tepid) has a ΔNPS of −48 — it is being read by the AI but not believed by it.

Negative ΔNPS means the AI is judging the brand more harshly than the human corpus does; the translation layer is leaving signal on the table. The corpus is favorable; the machine's representation of the corpus is not.

Something in the translation layer — the model's training data cutoff, the weighting of negative signals, the dominance of one bad piece of long-form content over fifty positive short-form ones, the live-retrieval layer pulling in a recent forum thread — may be suppressing the positive signal in the corpus and over-indexing on something less flattering. Identifying which axis is the source of the suppression is the diagnostic value of running both scores side by side.

6.1 The Aging Rubric

The λ-weighted NPS function uses an exponential decay kernel to age signals over time:

```
w_i(t) = e^(−λ · (t − t_i))
```

where t is evaluation time, t_i is the signal's timestamp, and λ is the decay rate (units of 1/time).

The intuitive calibration lever is half-life, not λ directly:

$$t_{1/2} = \ln(2) / \lambda$$

Common settings:

- **30-day half-life:** λ = 0.0231 per day (high sensitivity, noisier)
- **90-day half-life:** λ = 0.0077 per day (balanced)
- **180-day half-life:** λ = 0.0039 per day (stable, lagging)
- **365-day half-life:** λ = 0.0019 per day (long-memory)

Exponential decay is preferred over a hard cutoff: signals lose influence continuously without ever fully disappearing, avoiding discontinuity artifacts at the boundary.

Refinements: per-signal-type λ (promoter vs detractor half-lives may differ); source-trust coefficients multiplying w_i; volume floor below which the score returns "insufficient signal."

6.2 Interpreting ΔNPS

ΔNPS is the diagnostic that emerges from running both scores. Its sign and magnitude carry distinct meanings:

ΔNPS STATE	INTERPRETATION
Positive ΔNPS	Public corpus stronger than machine outputs
Negative ΔNPS	Machine outputs exceed observable corpus sentiment
Near-zero ΔNPS	Corpus and machine outputs broadly aligned

A wide negative ΔNPS (λNPS < μNPS) appears, in our operational experience, to be addressable — and is sometimes described as λ-suppression, the machine output reading weaker than the corpus warrants.

The interpretation is direct: the AI is judging the brand more harshly than the human corpus does, and the translation layer is leaving signal on the table. A wide positive ΔNPS (λNPS > μNPS) is rarer and usually points to either reputation laundering or to a corpus that is concentrated in low-weight social mentions while the machine is leaning on long-form sources we have not catalogued; this is the λ-amplification case, where machine output runs hotter than observable corpus sentiment, and may also indicate AI hallucination of positive signal or a corpus that under-represents real sentiment.

One honest scoping note. The λNPS pipeline is operational — it runs, it produces scores, it persists them, and the audit cost numbers above are measured from live runs. Calibration is open. By that we mean: we are not yet at the point of saying "a λNPS of +42 corresponds to such-and-such observed downstream business outcome." The relationship between λNPS and citation rate, between λNPS and traffic, between λNPS and conversion, is the next twelve months of work.

The claim today is that λNPS is a faithful, reproducible measurement of a real signal — the machine's revealed output — and that ΔNPS against μNPS appears to be a useful diagnostic. It is not yet a benchmark. We frame it as method, not as scoreboard.

6.3 Delivering at Machine Speed

The AI's reasoning budget is finite. When a model is composing an answer about a brand, it has a fixed window — measured in tokens, but bounded operationally by latency — in which to fetch and read the corpus material that will inform what it says. Time-to-first-byte and time-to-last-byte on the brand's published surface are not cosmetic performance metrics in this context. They directly determine how much of the brand's case the model gets to read before it has to commit to a response.

If the brand's site does not deliver content fast enough, stripped of unuseful characters like heavy html, the AI may cut off mid-read and synthesizes a response from whatever it managed to fetch. The picture is incomplete, the citation is weaker, and the machine fills the gaps with whatever the rest of the corpus told it — including the parts the brand did not write. A slow site is, functionally, a brand that does not get to make its full case in the room where the decision gets made.

This is the operational rationale for the translation layer. The translation layer's job is to deliver bot-formatted content — clean-room HTML, no JavaScript rendering, dense JSON-LD structured data, raw URLs as labels rather than anchor text, sub-200ms time-to-first-byte — so the AI gets the complete brand case in the time it has. The benchmarks we hold ourselves to are a p50 TTFB of ≤150ms from a pinned-IP, multi-hit canonical measurement, with TTLB as the headline KPI because TTLB is what bounds how much of the brand's case the model actually reads.

This is doctrinal, not optional. It is part of the translation-layer thesis, not a separate concept. A brand that has done the corpus work in §7 — long-form reviews, video testimonials, public forum responses, brand-authored counter-corpus on contested topics — but has not solved machine-speed delivery is shipping high-weight signal into a channel that cuts off before it gets read. The corpus side moves; the machine side does not move proportionately; ΔNPS stays wide. Closing ΔNPS requires the brand's published surface to deliver fully, fast, in the format the machine prefers, every time the machine asks.

7. Methodology Deep Dive

This section exists for the reader who wants to know exactly what the score does — every threshold, every weight, every fallback. The goal is reproducibility. If a brand's µNPS comes out at +34, that brand should be able to walk the score back to the individual signals that produced it, to the classifier prompt that read them, to the threshold that bucketed them, and to the version of the methodology that weighted them. The system is built around that traceability.

The score itself. µNPS is the weighted promoter percentage minus the weighted detractor percentage, multiplied by one hundred, rounded to one decimal:

```
weighted_total          = Σ signal.weight  (across all classified signals)
promoter_weighted_pct   = (Σ signal.weight where classification = "promoter")  /
weighted_total
detractor_weighted_pct  = (Σ signal.weight where classification = "detractor") /
weighted_total
µNPS_score              = (promoter_weighted_pct − detractor_weighted_pct) × 100
```

Weight provenance and versioning. The signal `weight` is set at ingest time, drawn from the `µNPS_weight_config` table by `(version, signal_type)`. Default v1 weights are the five-tier scheme in §3. Weight values are stored, not computed at score time, so a score's lineage is preserved even if the weight config is later edited. Every score row writes `methodology_version: 'v1'` (or v2, v3 in future).

Methodology versioning is not optional; it is the only thing that makes historical trend lines meaningful. Each score's `methodology_version` is the audit hook: a v1 score from October and a v2 score from February are not directly comparable headline-to-headline, and the schema makes that incomparability legible rather than hiding it.

Classification rubric. Each signal's `classification` is set by a Claude Haiku 4.5 call with prompt caching on the system rubric. The rubric instructs the model to return three fields:

- `sentiment_score`: a float from -1.0 (extremely negative) to +1.0 (extremely positive)
- `sentiment_label`: `positive` (score > 0.15), `neutral` (-0.15 to 0.15), or `negative` (< -0.15)
- `classification`: `promoter` (positive, advocacy intent), `passive` (neutral, functional language), or `detractor` (negative, complaint, dissatisfaction)

Audit and reproducibility. Every signal row carries `raw_text`, `source_url`, `keyword_boosts` (the matched keywords with + or - prefix), `intensity_modifiers`, and `raw_metadata` (everything the

source emitted, JSONB). A score from October can be opened in February and walked back to the exact 850 reviews that produced it, the exact classifications the model returned, and the exact weights applied. Nothing about the score is opaque. There is no proprietary black box between the corpus and the number; the box is glass.

> ***SIDEBAR: WHERE MNPS SITS IN THE ANSWERSHARE MEASUREMENT STACK.***
>
> *AnswerShare™ measures four axes of AI-mediated brand presence, and none of them substitute for the others.*
>
> ***ASQ™** (AnswerShare Score) measures what the brand publishes — content structure, retrievability, citation-worthiness, the quality of the surface AI bots read.*
>
> ***QFS** (Query Fan-Out Survivability) measures whether that content survives the fan-out queries an AI engine actually fires under the hood when answering a user.*
>
> ***Ghost-citation rate** measures the gap between citation and naming — when the machine draws on content without telling the user it came from a particular source.*
>
> ***µNPS / λNPS / ΔNPS** is the fourth axis, measuring what the corpus about the brand says, what the machines repeat from it, and the diagnostic gap between the two.*
>
> *ASQ is published surface. QFS is retrieval resilience. Ghost-citation is attribution leakage. µNPS / λNPS / ΔNPS is reputation. The four are independent measurements of related but distinct phenomena, and a complete read on a brand's AI-mediated standing requires all four.*

What the methodology refuses to do is also worth listing. It does not collapse promoter/passive/detractor into a single emotion gradient — the buckets are categorical, by design, and the bucketing happens at ±0.15. It does not infer signals the source did not emit. It does not impute scores for missing platforms; a brand with no Trustpilot presence simply has no Trustpilot signals, and the score reflects that. It does not adjust historical scores when methodology versions change. Each of those is a discipline that pulls in the same direction: when a brand's µNPS moves, the movement has to be the corpus moving, not the meter.

8. How to Affect It

> *You cannot rewrite what the internet has already said about you. You can add to it — and the additions that compound are the ones the model already trusts most.*

The first thing to say about moving the score is what cannot be done. The corpus is durable. Five-year-old Trustpilot reviews are still indexed, still readable, still weighted into the machine's view.

A brand cannot retroactively rewrite that material. It can dispute individual reviews; it can respond publicly to forum complaints; it can, in narrow cases of factual error, get something corrected. What it cannot do is replace the existing corpus with a different one. The corpus is what it is.

What a brand can do is *add* to the corpus, and the weights in §3 are explicit about which additions compound.

Long-form reviews (weight 1.5). A single 250-word Trustpilot review is weighted one-and-a-half times a short review in the methodology, and may be worth more than that in practice because the AI engines themselves appear to preferentially retrieve longer, fact-denser content.

Surfer's 289K-URL study (Kohli, Surfer SEO, 2026) found pages with ten or more discrete facts get cited at roughly twice the rate of fact-thin pages; the same dynamic appears to play out at the review level.

The work, for a brand, is creating natural occasions where satisfied customers can write at length. Post-purchase emails that ask one specific question ("what was the moment you knew you'd made the right decision?") rather than a generic five-star prompt tend to produce more long-form responses. The brand is not asking for the review; it is asking for the story, and the story is what the model reads.

Video testimonials (weight 2.0). The highest-credibility signal in the methodology, and the most expensive to produce. A video is hard to fake, hard to astroturf, and rich in retrievable detail — a face, a voice, an environment, often a recognizable product in frame. AI engines increasingly transcribe video at the retrieval layer; YouTube transcripts are first-class corpus material.

Ahrefs's brand-mention study reported YouTube mention correlation with AI recommendation at roughly **0.737 Spearman** (Ahrefs, *Brand Mentions and AI Recommendation Correlation Study*, 2025) — substantially higher than the all-channel 0.41 number — which is consistent with what the methodology already says: video is the highest-weight signal because the machine appears to treat it as such. Producing video testimonials is slow; it is also durable. A single well-shot customer story can compound across the corpus for years.

This means that the video transcript must be present when the page is read. To do that would make the page unreadable by humans. This is one feature of the AnswerShare solution.

Forum response patterns (weight 1.2). Reddit and BBB complaint threads are not the problem most brands think they are. A complaint thread with no brand response weighs as detractor signal at 1.2. A complaint thread with a public, substantive, named brand response *also* weighs as 1.2 — but the response itself becomes part of the long-form content the machine reads. The machine sees the complaint, and it sees the recovery.

Brands that respond publicly, by name, on the platform where the complaint was made, do not erase the detractor signal; they introduce a counter-signal the model weights alongside it. Silence, by contrast, leaves the detractor signal alone in the field. Solís's recommendation (Solís, *Humans of Martech*, 13 January 2026) that brands "align more closely with community managers" because community content is now first-class LLM source material is the operational form of this lever.

Forums are not the place reputation goes to die. They are the place reputation gets renegotiated, in public, in writing the machine will read.

Generic social mentions (weight 0.5). The lowest weight in the methodology, by design. A drive-by tweet, an Instagram caption, a TikTok comment — these are signals the machine sees but discounts. A brand spending its energy on social-mention volume is buying half-weight at scale, which is sometimes worth doing but is rarely the highest-leverage move. The math suggests redirecting effort to the higher-weight channels. A PR splash that does not land in review platforms, video, or substantive forum discussion may not move μNPS much.

The translation-layer lever. The fifth channel is the brand's own site — not the homepage, but authoritative content on the specific topics where the corpus disagrees with the brand. If the corpus claims a brand is slow on returns, and the brand has measurably fast return processing, the brand's own published, dated, structured content on return turnaround becomes a counter-corpus the machine can also retrieve. This is not "rewriting the corpus"; it is adding a high-credibility, brand-authored signal in a space the existing corpus is contested.

The AI engines do not weight this material the way they weight Trustpilot reviews — they shouldn't — but it is non-zero, and on contested topics it is the part of the corpus the brand controls. This is the part of the work that overlaps with ASQ: the published surface has to be structured, retrievable, and chunk-addressable for the machine to find it and use it. It is one lever among several, and it works only when the higher-weight channels are also being worked.

What this paper will not help with. This is the harder paragraph, and it goes here because the methodology is weaker without it. μNPS is not a metric a brand can game by producing fake testimonials, buying reviews, deploying astroturfed Reddit accounts, or hiring services that flood comparison sites with thin positive content. We say this not as a moral position — though we hold one — but as a measurement position.

The signals these tactics produce are, by design, the lowest-weight signals in the methodology: short, formulaic, low in specificity, low in intensity modifiers, often clustered by source or by time in ways that are detectable at the audit layer. They move the score the way water moves a sandbag. In the short run, possibly. In the medium run, the signal-quality dimensions of the methodology — `intensity_modifiers`, `has_specifics`, `length_chars`, the source-credibility weights — discount them aggressively.

In the long run, the machines themselves are getting better at the same detection: Trustpilot, Amazon, and the BBB have all invested heavily in fraud detection, review integrity, and trust-layer enforcement.

As AI systems increasingly shift toward grounded retrieval and trust-aware answer generation, platforms with stronger integrity controls appear more likely to be treated as reliable sources during citation and recommendation workflows.

A brand that pursues this strategy is paying for a number that does not survive its own ΔNPS. The corpus side moves slightly; the machine side does not move. The translation gap widens. The diagnostic gets worse. Astroturfed reviews are a way to fail the audit on both axes simultaneously.

The constructive version of the same point: the durable μNPS lift comes from the channels the model has structural reason to trust. Long-form, video, named-author, platform-verified, response-attached, brand-substantive. Those are the additions that compound. The rest is noise the methodology is calibrated to discount.

9. Limitations

This framework has substantial limitations. These include, but are not limited to:

Model nondeterminism. AI outputs vary across runs, prompts, temperatures, and retrieval states. λNPS reduces this through repeated sampling across the five canonical prompts and four engines, but variance does not go to zero.

Retrieval opacity. AI vendors do not disclose full retrieval or weighting systems. The methodology infers behavior from revealed outputs; it cannot inspect internal mechanics.

Correlation vs. causation. Observed correlations between mentions and recommendations — including the 0.41 Spearman (Kohli, Surfer SEO, 2026) and 0.737 Spearman (Ahrefs, 2025) figures cited above — do not establish causal relationships. They are consistent with a causal story; they do not prove one.

Prompt sensitivity. Small prompt wording changes may materially affect λNPS outcomes. The five canonical prompts are held constant by design, but a different prompt set could produce a meaningfully different score on the same brand.

Corpus survivorship bias. Public reviews overrepresent emotionally motivated users — the very satisfied and the very dissatisfied — at the expense of the silent middle. μNPS measures what got written, not what was felt by the unmeasured majority.

Platform weighting uncertainty. The true weighting systems used by AI vendors are unknown. The five signal-type weights in §3 are heuristic priors, informed by observed retrieval behavior and public research, not back-solved from vendor disclosure.

Language and regional bias. The current methodology is English-centric and may not generalize internationally without per-language calibration of both the classifier and the source-platform mix.

Review fraud and manipulation. Astroturfed reviews, coordinated attacks, and platform gaming remain ongoing challenges. The methodology's intensity modifiers and source-credibility weights discount thin or coordinated signals, but no detection layer is perfect.

Temporal lag. Training cutoffs and retrieval freshness differ substantially across engines. A corpus change made today will reach Perplexity (live-retrieval-heavy) sooner than it reaches a model relying primarily on a months-old training snapshot.

Methodology instability. Early-stage weights and thresholds remain experimental. Version 1 is a starting point, not a finalized standard, and v2 will almost certainly adjust some of these numbers.

The framework should therefore be interpreted as exploratory, operational, and iterative — not definitive.

10. Forecast

> *Your brand will be scored by machines whether you measure it or not. The brands that win the next decade are the ones that read their own score first and holistically optimize.*

We are reluctant, in this paper, to make quantitative forecasts. The systems being measured are not deterministic, the engines re-tune their weights without disclosure, and any number put on the page today will be wrong in three quarters of the predictions a year from now. What we will commit to is a directional read: λNPS — what the machines actually say about a brand — appears to be getting more consequential, not less, and the brands that measure it now, against the corpus that produced it, may have a structural advantage over the ones that wait.

The leading indicator we trust most operationally is ΔNPS. Across the brands we have audited so far, a wide negative ΔNPS (λNPS < μNPS) appears to be the most reliable early warning we have seen that AI-mediated brand search is suppressing a brand's real reputation. It suggests, before any traffic data confirms it, that the engines are reading a different version of the brand than the public has written.

Closing that delta — by adding high-weight signals where the corpus is contested, by responding publicly to forum complaints, by producing video testimonials in categories the machine currently weights lightly — appears, in our operational experience, to be addressable. It is slow. It is methodical. The evidence base is still small.

Three things to watch over the next twelve to twenty-four months.

Per-engine drift (λ-drift). A brand's λNPS on a given engine will likely drift as that engine retrains, re-weights, and re-tunes — what we call **λ-drift** when the machine sentiment for a given brand shifts measurably over time without a corresponding corpus change. We expect Gemini and ChatGPT to continue diverging on the mention-vs-citation axis Indig surfaced (Indig, Growth Memo, 2026). We expect Perplexity's live-search-heavy posture to make its λNPS more reactive to recent corpus changes than the others. We expect Claude's posture to remain conservative in the direction of caution — more passive classifications, fewer extreme positives or negatives, an internal preference for hedged language that may show up at the score level.

Tracking λ-drift week over week is not over-instrumentation; it appears to be the only way to see the engines moving against each other rather than as a single mass.

Industry baselines. The work no one has yet done at scale is establishing what "normal" looks like by sector. A direct-to-consumer mattress brand's μNPS distribution will not resemble a B2B SaaS infrastructure company's, and neither will resemble a consumer electronics retailer's. Without sector baselines, an individual brand's +27 is uninterpretable. Building those baselines — at least at the major-industry level — is part of the year ahead, and it is a prerequisite for any brand to know whether its score is good, average, or quietly bleeding.

The λNPS-to-citation-rate relationship. The most interesting open question is the order in which the signals move: corpus first, then machine, then citation — or some other sequence. We do not have a confident read yet. The working hypothesis is that μNPS leads as the substrate, λNPS follows after some lag determined by engine retraining cadence and live-retrieval weighting, and citation rate lags λNPS because citation is a downstream artifact of the machine's revealed output.

If that hypothesis holds, λNPS — read against its μNPS substrate, with ΔNPS as the diagnostic — may become a measurable forward indicator on AI-mediated brand outcomes, earlier than traffic, conversion, or any dashboard that currently exists. That is the prize. It is also unproven. The next twelve months will tell us.

What remains unknown is magnitude. The relationship between corpus sentiment, machine sentiment, citations, recommendations, traffic, and conversion has not yet been empirically mapped with confidence. Future work includes inter-rater agreement on the classification rubric, prompt-stability analysis across runs, longitudinal consistency tests, and correlation studies against external benchmarks. This paper should therefore be interpreted as a framework proposal rather than a finalized standard.

The forecast we will commit to is the discursive one. Every brand will likely have a λNPS the way every brand has a credit rating — a live, running score of what the machines are presently saying about it, computed whether the brand asks for it or not. The brands that read it first — that measure λNPS against the μNPS substrate, that close ΔNPS before it becomes a citation deficit, that treat AI-mediated brand presence as a channel to be measured rather than hoped about — appear to have a years-long lead on the ones that wait. The work is not glamorous. The signals are not loud. The compound is the point.

Suspected Manipulation Impact on AI Trust

One of the most significant emerging risks in AI-era visibility may be the long-term trust impact associated with manipulative publishing behavior. Traditional search engines frequently relied on explicit ranking penalties to combat tactics such as keyword stuffing, doorway pages, hidden text, and link manipulation. Modern AI systems may behave differently.

Rather than issuing discrete penalties, AI retrieval and grounding systems may instead develop probabilistic trust suppression behaviors around domains exhibiting patterns associated with manipulation or adversarial optimization. This distinction is important. A search engine penalty is often explicit and recoverable through corrective actions. AI trust degradation may instead emerge gradually through accumulated behavioral signals, making recovery slower, less predictable, and more difficult to diagnose.

Large language models and retrieval systems are particularly sensitive to manipulation risk because generated answers inherit reputational liability from the sources they incorporate. Hallucinations, misinformation, and low-confidence outputs directly damage trust in the AI system itself. As a result, AI systems are structurally incentivized to avoid sources that appear engineered primarily to exploit retrieval mechanics rather than provide reliable information.

Modern AI systems are likely capable of detecting many forms of unnatural optimization behavior more effectively than traditional search engines. Patterns such as repetitive keyword density, semantically awkward phrasing, excessive template duplication, low-information-density content, synthetic engagement patterns, contradictory claims, or over-engineered citation targeting may collectively function as indicators of adversarial intent.

Importantly, the issue may not be manipulation alone, but the inference of manipulation intent. AI systems increasingly appear optimized to identify whether a source behaves like a trustworthy publisher. Domains associated with persistent manipulation signals may experience:

- reduced retrieval priority
- lower reranking confidence
- increased corroboration requirements
- reduced grounding confidence
- citation reluctance
- answer exclusion during high-confidence generation

These effects may compound over time as retrieval systems accumulate historical quality signals. In this sense, AI trust may behave more like reputation or creditworthiness than traditional SEO ranking. Once trust degradation occurs, rebuilding confidence may require prolonged periods of semantic consistency, grounded publishing behavior, corroboration from trusted sources, and sustained reduction of adversarial signals.

As AI-generated answers increasingly become the dominant discovery interface, this dynamic may fundamentally alter digital publishing incentives. In the search era, manipulation often sought visibility. In the AI era, manipulation may directly erode trust itself.

11. Conclusion

AI systems increasingly generate responses that users interpret as reputational judgments. Those judgments are the live brand surface — what the next ten thousand users hear — shaped, at least in part, by public corpus signals the brand does not control.

λNPS operationalizes measurement of that live surface: what the machines currently generate about a brand under controlled prompting. μNPS measures the corpus substrate underneath. ΔNPS, the gap between the two, is the diagnostic that emerges when both run side by side. The framework emphasizes reproducibility, methodology versioning, transparency about weights and thresholds, and explicit acknowledgment of where claims are operational rather than proven.

AI doesn't rank you. It forms an opinion of you. Then it tells the next 10,000 users what it thinks. λNPS is how a brand reads that opinion as it is currently expressed. μNPS is how a brand reads the corpus underneath. ΔNPS is the gap between them — and the gap is increasingly where brand work happens.

The goal is not to claim certainty about AI internals. The goal is to propose a measurable framework that the community can test, challenge, refine, or reject based on evidence. That process is the point.

References

- Muller, Britney. Interview, *AI SEO Show* (YouTube), 1 November 2025. https://www.youtube.com/watch?v=l4fIHPtjIMY
- Muller, Britney. Interview, *Elumynt UpArrow podcast*, 2025. https://www.elumynt.com/podcast/actionable-ai-for-marketers-the-human-in-the-loop-with-britney-muller
- Kohli, Saloni. *SEO Trends 2026*. Surfer SEO, 8 May 2026. https://surferseo.com/blog/seo-trends-2026/
- Indig, Kevin. The Ghost Citation Problem. Growth Memo, 20 April 2026. https://www.growth-memo.com/p/the-ghost-citation-problem
- Solís, Aleyda. Interview, *Humans of Martech* podcast, episode 202, 13 January 2026. https://humansofmartech.com/2026/01/13/202-aleyda-solis-ai-search-crawlability/
- Solís, Aleyda. *The 10 Steps AI Search Content Optimization Checklist*. aleydasolis.com, 25 July 2025. https://www.aleydasolis.com/en/ai-search/ai-search-optimization-checklist/
- Ahrefs. *Brand Mentions and AI Recommendation Correlation Study*, 2025. (YouTube mention 0.737 Spearman correlation with AI recommendation.)
- Google Cloud. How Vertex AI Grounding helps build more reliable models. Google Cloud Blog. https://cloud.google.com/blog/products/ai-machine-learning/how-vertex-ai-grounding-helps-build-more-reliable-models
- Anthropic. Introducing Contextual Retrieval. anthropic.com/engineering. https://www.anthropic.com/engineering/contextual-retrieval
- Song, Maojia, et al. Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse. arXiv:2409.11242, 17 September 2024. https://arxiv.org/abs/2409.11242

Acknowledgment

Portions of this paper were developed with the assistance of generative AI tools used for drafting, editing, synthesis, and analytical exploration.