Barkhausen AI

CONVENTION · BA-C-3

Version 1.0.1 · Stable

Sampling-protocol requirements

AI assistants return different answers to the same question from one request to the next, so any measurement of how visible a brand is inside those answers is a statistical estimate, not a reading.

“Every leap begins as a crackle.”

DOCUMENT	BA-C-3
VERSION	1.0.1 (Stable)
EFFECTIVE	2026
SUBJECTS	Measurement & statistics; Conventions & policy
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

AI assistants return different answers to the same question from one request to the next, so any measurement of how visible a brand is inside those answers is a statistical estimate, not a reading. This convention states the requirements a sampling design must meet for its estimates to be credible. It summarizes the public evidence that a single observation is uninformative and derives the sample size a stated confidence interval requires. It requires interval methods that stay valid near zero and one, representation of each query as a distribution over real phrasings, partial pooling, observations spread across the stated window, engine-version and measurement-channel disclosure, controlled region and personalization, change-point monitoring for silent engine updates, multiplicity disclosure, and refusals recorded as availability observations. It formalizes the Barkhausen Criterion — the conditions under which a claimed visibility change counts as real — and closes with a reporting checklist.

CONTENTS

A measurement of AI visibility estimates how often an assistant mentions a given brand or entity when asked about a topic. Because the answer an assistant returns to the same question changes from one request to the next, such a measurement is a statistical estimate, not a reading off an instrument. This convention states the requirements a sampling design must satisfy for its estimates to be credible. It is written at the level of requirements — what a valid design must be true of — so that a reader can judge whether a study's conclusions follow from its data. It does not describe any particular measurement system or tool.

The quantity being estimated must be stated precisely before any sampling requirement makes sense. For a brand or entity B, an information need k — a topic query, represented by its phrasing distribution rather than one sentence — an engine e, a time window t, and a region g, the estimand is a probability:

$\theta(B, k, e, t, g) = P(\text{answer mentions } B \mid \text{prompt} \sim Q_k,\ e,\ t,\ g).$

Here Q_k is the distribution of real user phrasings for query k — the set of semantically equivalent ways people actually ask it — not a single sentence. This estimand is the **Visibility Probability** (**VP**). Every requirement below is a condition under which an estimate $\hat{\theta}$ of this quantity is trustworthy.

Normative note. The words MUST, MUST NOT, SHOULD, SHOULD NOT, and MAY carry their established normative meanings. MUST is an absolute requirement; a design that violates a MUST does not produce a credible visibility estimate. SHOULD is a strong recommendation that may be set aside only with a stated, defensible reason recorded in the study. MAY marks an option that is permitted but not required. Requirements are stated so that compliance can be checked from a study's disclosure alone, without access to its raw data.

1 The instability of a single observation

The foundation of every requirement that follows is a single empirical fact: one answer to one query is not a usable estimate of visibility. Three independent lines of public evidence establish this, at different time scales.

Prompt-rewrite sensitivity. A 2026 study of production retrieval-augmented recommendation systems ran roughly 12,000 queries across deployed OpenAI and Anthropic models, comparing a same-prompt rerun baseline against natural-language paraphrases of the same underlying need [1]. Where repeating an identical prompt reproduced the recommendation set with a Jaccard similarity of about 0.50 to 0.61 within an engine, merely cosmetic rewordings of the same need cut that similarity to about 0.29 — 21 to 32 percentage points below the rerun baseline — and rewordings that added a qualifying constraint cut it to about 0.14 — a level the study places 37 to 48 percentage points below the rerun baseline. Two conclusions follow. First, a change in wording that preserves the underlying need can still replace much of the recommendation set, so a single fixed sentence is a fragile stand-in for a query. Second, the effect held across the deployments tested and was not removed by increasing reasoning effort, so instability must be measured rather than assumed away.

Day-to-day drift. A 2026 study tracked repeated queries across four AI answer engines over a window of about a month and a half and found that identical prompts issued on two consecutive days produced answers whose cited-source sets overlapped only about 34% to 42%, and whose sets of mentioned brands overlapped about 45% to 59% — the brand figure computed on the three of the study's four verticals that met its brand-detection threshold [2]. Repeated runs of the same prompt within a single day showed source overlap in the same range (about 32% to 43%), so much of the variation is request-to-request stochasticity rather than genuine day-over-day movement. The same question, unchanged, yields substantially different sources and a materially different roster of brands. Brand mentions are somewhat

more stable than the underlying citations, which is why a visibility estimate defined on brand mentions is a firmer basis for measurement than one defined on cited sources — but neither is stable enough for a single reading.

Slower change, and differences between engines. Beyond the day scale, engines are periodically rebuilt outright: models are replaced, retrieval systems are reconfigured, and system prompts change, all without announcement, so drift compounds over months. No independently verified figure for monthly citation turnover exists in the public record, and this convention treats the monthly scale qualitatively. What the public evidence does document is that instability differs sharply between engines: measured within a single 24-hour span, per-engine source overlap averages about 0.23 on one engine and about 0.51 on another [2]. A stability figure pooled across engines therefore describes none of them, which is why the reporting requirements below are stated per engine.

These findings describe variation at three time scales, and a sound design treats them differently. Fast, request-to-request variation — decoding randomness and retrieval variation — is observation noise, to be averaged out by repeated sampling. Gradual turnover over weeks and months is slow drift, and is the legitimate target of a trend estimate. An abrupt, system-wide shift is a structural break, typically an engine update, to be detected rather than absorbed into a trend. Three further factors — phrasing, region, and personalization — are not noise at all but design variables that a protocol must control or stratify on. A single observation collapses all of these into one number and can distinguish none of them; that is why θ cannot be estimated from a single draw. Figure 1 places these three scales on one time axis, with the sampling response each calls for.

The consequence is quantitative, not merely cautionary. Suppose a query's true visibility is $p = 0.5$. A single answer is one Bernoulli draw with standard deviation $\sqrt{p(1-p)} = 0.5$ — the estimate sits at the noise floor of the measurement. Two single observations compared across periods can produce an apparent movement that looks like several standard errors under an assumption of independent, identically distributed sampling, yet that assumption is falsified the moment phrasing, interface, or time-of-day effects are present, and all three are present in practice. An apparent change built on single observations therefore cannot be attributed to any intervention. A protocol that reports movement from single-shot queries reports noise.

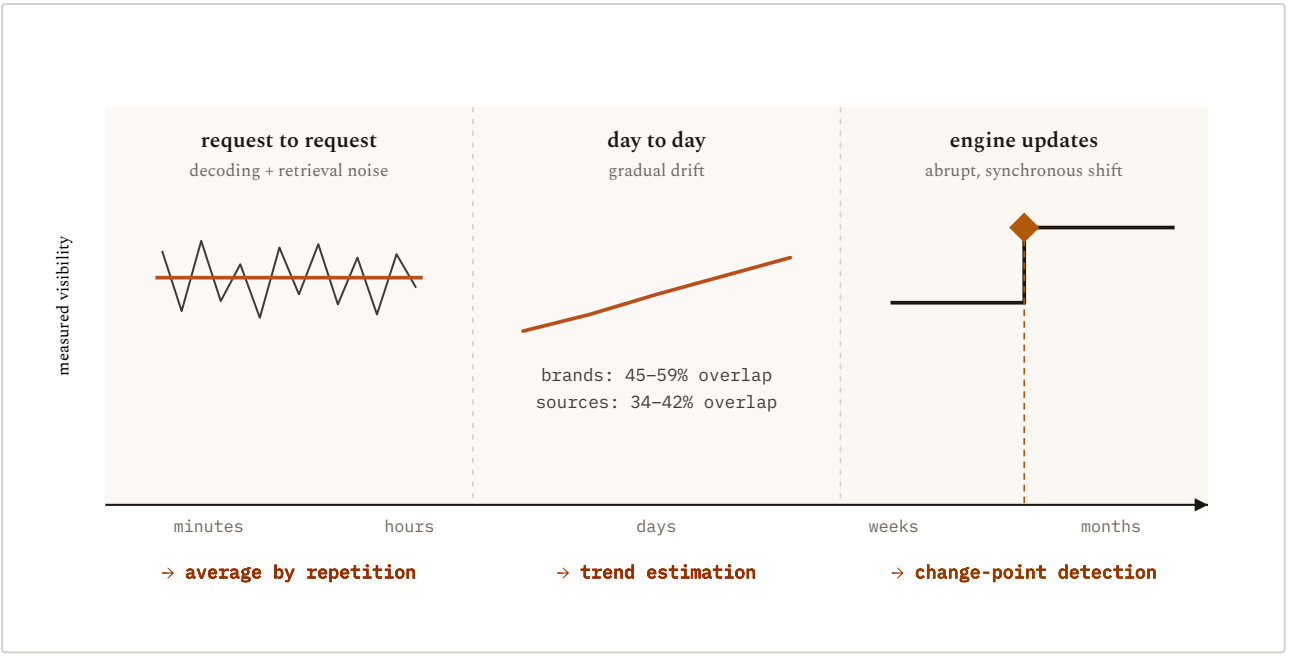

Figure 1. Three timescales of variation, three responses. Request-to-request noise — decoding and retrieval variation — is averaged out by repeated sampling; gradual day-to-day drift, over which identical prompts on consecutive days shared only 45–59% of mentioned brands and 34–42% of cited sources, is the target of trend estimation; and an abrupt, synchronous shift across many queries, the signature of a silent engine update, is isolated by change-point detection rather than absorbed into a trend.

Barkhausen AI · BA-C-3; overlap figures from arXiv:2604.07585 [2]

2 Sample size

Because $\hat{\theta}$ is a proportion, the sample size needed to pin it down follows from the Bernoulli model. With n independent observations of which k mention the brand, the point estimate is $\hat{p} = k/n$, its standard error is $\sqrt{\hat{p}(1-\hat{p})/n}$, and the half-width of a 95% normal-approximation confidence interval is $E \approx 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. Solving for n:

$$n \approx \frac{1.96^2\, p(1-p)}{E^2}.$$

The variance $p(1-p)$ is largest at $p = 0.5$, so planning at $p = 0.5$ gives the worst-case, conservative sample size for a target interval half-width E.

TARGET INTERVAL HALF-WIDTH E (AT $P = 0.5$)	REQUIRED N
$\pm 10\%$	≈ 96
$\pm 5\%$	≈ 385

A protocol MUST state a target interval width and MUST justify its per-cell sample size n against that target. A "cell" here is one combination of query, engine, measurement channel, window, and region (channels are defined below); the sample size requirement applies within each cell, because an estimate is only ever an estimate of θ for a specific cell. Reporting a visibility percentage without a companion sample size and interval is non-compliant: the table above shows that a $\pm 5\%$ claim demands four times the sample of a $\pm 10\%$ claim — and either demands roughly two orders of magnitude more than the single observation many published claims rest on — while a bare percentage hides which one, if any, was actually achieved.

A protocol MAY allocate sampling adaptively rather than fixing n in advance — for example, stopping early on queries whose visibility is unambiguously high or low and concentrating observations on queries near a decision threshold — provided the stopping rule controls error rates and the achieved n and

interval are reported per cell. The classical machinery for error-controlled sequential stopping exists and is well understood [9][10]; adaptive allocation changes how observations are distributed across queries, not what the final estimate must disclose.

3 Interval estimation

The normal-approximation interval used to derive the sample size above is itself unreliable at the edges of the scale. When $\hat{p}$ is near 0 or near 1, the symmetric interval $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ can extend below 0 or above 1, and its coverage — the probability that it actually contains the true θ — falls well short of the nominal 95%. A visibility of 5% with a small sample can be assigned a normal interval with a negative lower bound, which is meaningless.

Where an estimate lies near 0 or 1, a protocol MUST use an interval method that stays valid there. The **Wilson score interval** [3] and the **Clopper–Pearson exact interval** [4] both keep their bounds inside $[0, 1]$ and hold their coverage in this regime; standard statistical libraries implement both. The Wilson interval is the lighter default and has good coverage across the full range; Clopper–Pearson is conservative and guarantees at least nominal coverage. Away from the edges the three methods nearly coincide, so a protocol that uses Wilson or Clopper–Pearson throughout loses nothing and is safe everywhere. What is not acceptable is reporting a near-zero or near-one visibility with a naive normal interval.

4 Representing a query as a phrasing distribution

The estimand θ is defined over Q_k, the distribution of real phrasings for query k. This is not a modeling nicety; it is forced by the rewrite-sensitivity evidence. Representing a query by one fixed sentence assumes that Q_k is a degenerate distribution — all its mass on a single phrasing — and the finding that paraphrases preserving the same underlying need cut recommendation-set overlap far below the same-prompt rerun baseline [1] falsifies that assumption directly. A single-sentence measurement estimates θ at one arbitrary point of Q_k and silently reports it as if it were the whole query.

A protocol therefore MUST represent each query by a set of real, semantically equivalent phrasings drawn from how users actually ask, not by one sentence. The phrasings SHOULD cover the genuine variation in user wording — length, directness, and, where the audience is multilingual or multi-regional, language and locale. A protocol MUST disclose the number of distinct phrasings used per query and describe, in general terms, how the range of user wording was covered. This disclosure is at the level a reviewer needs to judge whether the query was represented fairly; it is not a requirement to publish the phrasings themselves, and the specific phrasings are not the object of this convention.

Two failure modes are worth naming. A single-phrasing design understates uncertainty, because it never exposes the between-phrasing variation that the evidence shows to be large. And a design that uses several phrasings but conceals how many, or how they were chosen, cannot be assessed for coverage — a handful of near-identical rewordings is not a representation of Q_k, and only disclosure of the count and the coverage strategy lets a reader tell the difference.

5 Pooling across phrasings

Once a query is sampled across multiple phrasings, the estimates must be combined into a single query-level $\hat{\theta}$. The naive approach — estimate each phrasing independently and average, or report each on its own — wastes information and misleads, because phrasings with few observations carry large sampling

noise that a simple average passes straight through. A phrasing sampled ten times can show 0% or 100% by chance alone, and treating that as its visibility distorts the query estimate.

The correct treatment is a hierarchical, partial-pooling estimator [7][8]. Phrasings are modeled as belonging to a query, and each phrasing's estimate is shrunk toward the query-level mean by an amount governed by its own sample size: a well-sampled phrasing keeps most of its own signal, while a sparsely sampled one borrows strength from the query as a whole and is pulled toward the common mean. This is the standard remedy for exactly this structure — many related proportions, unevenly sampled — and it produces a query-level posterior mean with a credible interval that honestly reflects both within- and between-phrasing variation. Conceptually it can be realized with a conjugate Beta–binomial model at the query level or, where a fuller query-phrasing-time structure is wanted, a multilevel logistic model; the requirement is the shrinkage behavior, not any particular implementation.

A protocol that estimates a query's visibility from multiple phrasings SHOULD use partial pooling rather than reporting phrasings independently or giving each equal weight regardless of its sample size. Where partial pooling is not used, the protocol SHOULD state why and SHOULD show that the reported interval still accounts for between-phrasing variation.

6 Time windows, repetition, and engine disclosure

An estimate of θ is inseparable from the window over which it was collected, because visibility drifts. An estimate MUST state its time window — the dates over which its observations were gathered. A percentage attached to no window is not interpretable: the day-to-day and month-to-month evidence above shows that the same query's visibility is a moving quantity, so a number without a window names no quantity at all.

Within a window, sampling MUST be distributed across the window rather than collected in one burst: an estimate whose stated window is a week but whose observations were all gathered in one hour describes that hour, not that week. Where the window spans multiple days, sampling SHOULD cover at least three distinct days. Repetition within the window is what converts a set of unstable single answers into a stable estimate of the window's central tendency. The window should be short enough that the underlying θ is roughly constant across it, and long enough to accommodate the repetition the sample-size target demands.

Observations collected close together also share retrieval state — the same index snapshot, the same load conditions — and are therefore positively correlated. Correlated draws carry less information than independent ones: the *effective sample size* is smaller than the raw count, and an interval computed as if the draws were independent is too narrow. A protocol SHOULD account for this clustering, by spacing observations in time, by reporting an effective-sample-size estimate alongside the raw n, or by using an interval method that models the correlation — the hierarchical estimator of the pooling section extends naturally to a query–phrasing–day structure [8].

Every estimate MUST name the engine and its version or observation date, and estimates MUST be reported per engine: because instability itself differs sharply between engines, a cross-engine aggregate MAY be published only alongside the per-engine figures it was built from. Results are perishable, and the honesty boundary must be stated plainly: every estimate describes the engine as sampled during its stated window; engines change without notice, and results should be assumed perishable. An estimate that names no version cannot be reproduced, compared across time, or defended when an engine changes, because there is no record of which engine produced it.

7 Region and personalization

Retrieval and ranking differ by region, so region is a design variable, not a nuisance. An estimate MUST state the region it describes. A visibility figure that pools or leaves unstated the regions it was collected from confounds genuine cross-region differences with sampling variation and cannot be interpreted as the visibility in any one market.

Personalization and conversational memory are likewise design variables: a logged-in session's history and stored memory change the answers an engine returns, so an uncontrolled account measures the interaction between a query and one account's history rather than the query itself. A protocol MUST disable or otherwise control personalization and memory — for example by using stateless sessions that accumulate no history, or accounts configured to carry no memory across observations — and MUST state which. The requirement is that the measured quantity is the query's visibility, not an artifact of one account's past.

8 Measurement channels

The same engine is usually reachable through more than one deployment surface: the consumer interface that users actually see, one or more official APIs, and third-party intermediaries that resell access. These are different deployments of e, not different views of one deployment. The metric convention already fixes the interface as part of the engine's identity (BA-C-2 §1); this section states the sampling consequences. Retrieval configuration, model routing, tool invocation, and answer composition can all differ between surfaces, so an estimate obtained through one channel is an estimate *for that channel* — a brand can be absent from an engine's consumer answers while present in the same engine's API answers, or the reverse, within the same window.

Three requirements follow. A protocol MUST disclose the measurement channel of every estimate — consumer interface, official API, or other, with the interface or endpoint class named — as part of the cell definition. A claim about what *users* see MUST either be measured on the consumer interface or be accompanied by a stated comparison of the collection channel against consumer-interface samples for the same cells and window; where such a calibration is used, the protocol SHOULD quantify the observed divergence rather than assert equivalence. And observations from different channels MUST NOT be pooled into one estimate without disclosure; pooled cross-channel figures inherit the worst interpretability of both sources.

9 Detecting engine change

Engines are updated silently: providers change weights, retrieval strategies, and system prompts without announcement. Such a change is exogenous to any brand's actual standing, and it typically shows up the same way — a large, same-direction shift across many unrelated queries on the same day. If it is read as real movement, every brand in the study appears to move at once, and the measurement misattributes an engine update to the brands it is tracking.

A monitoring design SHOULD run online change-point detection on the per-query visibility series so that a large simultaneous shift across many queries is flagged as a likely engine change rather than reported as genuine movement [5][6]. The distinguishing signature is breadth and synchrony: an isolated move on one query is ordinary drift or noise, while a coordinated jump across a large fraction of queries on a single date is the fingerprint of an engine change and SHOULD be surfaced as such.

One further hazard is built into monitoring itself. A design that tracks many cells — tens of queries across several engines and windows — performs many implicit significance tests, and some cells will cross any fixed threshold by chance alone. A protocol MUST disclose the number of cells monitored alongside any claim that particular cells changed, and a study that reports which cells changed SHOULD apply a multiple-comparison control (a false-discovery-rate procedure is the natural fit for monitoring panels) or state explicitly that none was applied.

10 Deciding that a change is real: the Barkhausen Criterion

Continuous optimization effort produces visibility change that arrives, when it arrives, as discrete jumps against a noisy, drifting background. The question a measurement must answer is whether an observed jump is real — attributable to a durable shift in the engine's treatment of the entity — or an artifact of sampling noise, drift, or an engine update. This convention names the decision rule for that question the **Barkhausen Criterion**. A claimed visibility change passes the Criterion only if all four of the following hold:

1. **Significance against proper intervals.** The change is established by comparing interval estimates obtained under this convention — a two-proportion comparison, or a model-based contrast from the hierarchical estimator of the pooling section — at a stated confidence level. A difference between two point estimates whose intervals were never computed does not qualify.
2. **Sustainment.** The shifted level persists for at least K consecutive sampling windows, with $K \geq 2$ and both K and the window length disclosed. A jump observed in a single window is a candidate, not a change.
3. **Engine-change exclusion.** The jump does not coincide with a flagged engine-change point (the broad, synchronous shift described above). Where it does, the claim MUST be re-based on windows after the engine change, because the movement is otherwise attributable to the platform rather than the entity.
4. **Multiplicity honesty.** The claim discloses how many cells were monitored, and the significance treatment accounts for that number as required above.

A change claim that satisfies all four is a *visibility jump* in the sense this publication uses the term. A claim that fails any one of them is reported, if at all, as an observation awaiting confirmation. The Criterion is deliberately stated as requirements on disclosure and design rather than as a single test statistic: implementations may reasonably differ in the test they apply, but a reader must be able to verify each of the four conditions from the study's disclosure alone. Change-point monitoring is the mechanism that makes the third condition checkable rather than asserted.

11 Refusals and availability

An engine sometimes declines to answer — because a query touches a restricted topic, trips a safety filter, or is otherwise refused. A refusal is not a failed observation to be dropped; it is an observation about the query, specifically about its availability. Discarding refusals as missing data biases the visibility estimate, because it silently conditions on the engine having been willing to answer, which is not the population the study means to describe.

A protocol MUST record refusals as observations and MUST report the refusal rate per cell alongside the visibility estimate. The refusal rate is itself a result — a query that is frequently refused is, for practical purposes, one on which the brand cannot be visible regardless of any optimization — and it belongs in the record next to $\hat{\theta}$, not in the gap left by deleting it. Where refusals are frequent, the visibility estimate

SHOULD be reported as conditional on a non-refused answer, with the refusal rate stated so that the unconditional availability of the query is legible.

12 Reporting checklist

A study whose sampling design complies with this convention discloses, for each visibility figure it publishes:

1. **The estimand.** The brand or entity, the query, the engine and version or observation date, the time window, and the region the figure describes.
2. **The measurement channel.** The deployment surface the observations came from — consumer interface, official API, or other — and, for claims about what users see, the consumer-interface basis or the stated cross-channel calibration.
3. **The point estimate with an interval.** $\hat{\theta}$ reported with a confidence or credible interval and the per-cell sample size n — never a bare percentage.
4. **The target and the achieved precision.** The target interval width the design was sized for, and the n that meets it, justified against the Bernoulli sample-size relation.
5. **The interval method.** Wilson score or Clopper–Pearson where any estimate lies near 0 or 1.
6. **The phrasing representation.** The number of distinct phrasings per query and a general description of how user-wording variation was covered.
7. **The pooling method.** Whether and how partial pooling was applied across phrasings, or a stated reason and an alternative that still accounts for between-phrasing variation.
8. **Distribution across the window.** That sampling was spread across the stated window rather than taken in one burst, and how within-window clustering was handled (spacing, an effective-sample-size estimate, or a correlation-aware interval).
9. **Per-engine reporting.** Estimates stated per engine; any cross-engine aggregate published only alongside its per-engine components.
10. **Region and personalization controls.** The region, and the statement that personalization and memory were disabled or controlled, with the method named.
11. **Change monitoring and multiplicity.** For a monitoring study, that a change-point method flags system-wide shifts as likely engine changes; the number of cells monitored; and the multiple-comparison treatment (or its stated absence).
12. **Change claims.** Any claim that visibility changed satisfies the four conditions of the Barkhausen Criterion, with K and the window length disclosed.
13. **Refusals.** The refusal rate per cell, recorded as an observation, with the visibility estimate marked conditional where refusals are frequent.
14. **The perishability boundary.** The statement that results describe the engine as sampled during the stated window and should be assumed perishable.

A study that satisfies all fourteen is one whose visibility estimates can be judged, reproduced within the limits of a changing engine, and defended.

13 Limitations

These requirements bound what a credible measurement must satisfy; they do not by themselves guarantee a correct one. Several limits should be stated honestly.

The primary rewrite-sensitivity evidence comes from a single domain — production retrieval-augmented commercial recommendation — and two model families [1]. It measures paraphrase effects against a same-prompt rerun baseline and finds them well beyond run-to-run randomness, but the magnitude it reports should not be assumed to carry over unchanged to other domains or engines. The day-to-day and between-engine evidence was collected on a different, small set of engines over a single multi-week window [2]. Instability is engine-, domain-, and window-dependent, and the specific figures cited here describe the systems studied, not answer engines in general. For change on the scale of months, no independently verified public figure exists; this convention makes no quantitative claim at that scale.

The Barkhausen Criterion is stated as four checkable disclosure conditions, not as a single prescribed test. Two studies can both satisfy it while using different significance machinery, and the Criterion does not by itself guarantee that a sustained, significant, engine-change-free jump was caused by any particular intervention — causal attribution requires design beyond the scope of this convention.

The sample-size relation assumes independent, identically distributed observations within a cell. Real sampling departs from this: observations within a short window are not fully independent, and stratifying across phrasings introduces structure the simple formula ignores. The formula therefore gives a planning target and a floor, not an exact requirement; the partial-pooling and interval requirements exist precisely because the i.i.d. proportion is an approximation. Reported intervals are conditional on the model used to compute them, and a mis-specified pooling model can produce intervals that are too narrow.

Finally, this convention governs sampling design, not the definition of a "mention" or the accuracy of the extractor that detects one. Two studies that both comply here can still disagree if they count mentions differently or if their extractors have different error rates. The perishability boundary is not a limitation to be engineered away but a permanent property of the object being measured: engines change without notice, and every figure a compliant study publishes describes an engine that may already have moved on.

REFERENCES

1. Jack, Lehman, Maloney, Xu; arXiv:2605.27440. *Paraphrase Brittleness in Production Retrieval-Augmented Commercial Recommendation: Reproducibility Below the Rerun-Stability Baseline* (2026). https://arxiv.org/abs/2605.27440 Accessed 2026-07-08. [archived]
2. Schulte, Bleeker, Kaufmann; arXiv:2604.07585. *Don't Measure Once: Measuring Visibility in AI Search (GEO)* (2026). https://arxiv.org/abs/2604.07585 Accessed 2026-07-08. [archived]
3. E. B. Wilson; Journal of the American Statistical Association 22(158), 209–212. *Probable inference, the law of succession, and statistical inference* (1927). https://doi.org/10.1080/01621459.1927.10502953 Accessed 2026-07-08.
4. C. J. Clopper and E. S. Pearson; Biometrika 26(4), 404–413. *The use of confidence or fiducial limits illustrated in the case of the binomial* (1934). https://doi.org/10.1093/biomet/26.4.404 Accessed 2026-07-08.
5. E. S. Page; Biometrika 41(1-2), 100–115. *Continuous inspection schemes* (1954). https://doi.org/10.1093/biomet/41.1-2.100 Accessed 2026-07-08.
6. R. P. Adams and D. J. C. MacKay; arXiv:0710.3742. *Bayesian online changepoint detection* (2007). https://arxiv.org/abs/0710.3742 Accessed 2026-07-08. [archived]
7. B. Efron and C. Morris; Journal of the American Statistical Association 70(350), 311–319. *Data analysis using Stein's estimator and its generalizations* (1975). https://doi.org/10.1080/01621459.1975.10479864 Accessed 2026-07-08.
8. A. Gelman and J. Hill; Cambridge University Press. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (2007).
9. A. Wald; The Annals of Mathematical Statistics 16(2), 117–186. *Sequential Tests of Statistical Hypotheses* (1945). https://doi.org/10.1214/aoms/1177731118 Accessed 2026-07-09.
10. K. K. G. Lan and D. L. DeMets; Biometrika 70(3), 659–663. *Discrete sequential boundaries for clinical trials* (1983). https://doi.org/10.1093/biomet/70.3.659 Accessed 2026-07-09.

HOW TO CITE

Barkhausen AI (2026). Sampling-protocol requirements. https://barkhausen.ai/conventions/sampling-requirements/

Barkhausen AI · Est. MMXXVI · *Every leap begins as a crackle.*