XPOSER.AI

BENCHMARK · METHODOLOGY EDITION

The 60-Second Triage Benchmark.

The definitions, measurement protocol and publication rules for our monthly report on triage latency and AI/human classification agreement — published before the first numbers, so the numbers can be judged against a method nobody chose after seeing them.

Published	September 2026
Edition	Methodology 1.0
Supersedes	Nothing — this is the first edition
Audience	Compliance leaders, internal audit, procurement, model risk
Author	XPOSER.AI

What this edition contains, and what it does not. This document specifies *how* the benchmark is measured. It deliberately contains no performance figures. Publishing the method first is the point: a metric definition written after the results are known is a metric definition chosen to flatter them. Section 9 states when measured editions publish and what each will contain.

CONTENTS

Section 01

Why A Vendor Should Publish Its Method First

Every vendor in this category quotes an accuracy figure. Almost none publishes the definition that produced it, which is what makes the figure meaningless.

“94% accurate” is unfalsifiable without four missing pieces of information: what was being classified, into how many categories, against whose judgement, and on what sample. Change any one and the number moves by tens of points. A vendor free to choose those four after measuring can produce almost any figure it likes without stating a single falsehood.

The September 2024 Evaluation of Corporate Compliance Programs asks a question that makes this a compliance problem rather than a marketing one:

> *“What baseline of human decision-making is used to assess AI?”*
>
> U.S. DOJ, Evaluation of Corporate Compliance Programs, September 2024

A customer who cannot answer that about their own triage system has an evidence gap. A vendor who cannot supply the underlying method has caused it. This document exists so that our customers can answer the question with a citation rather than an assurance.

The commitment this edition makes

The definitions in sections 3 to 6 are fixed before any measured edition is published. If we later change a definition, section 9 requires us to say so in the edition where the change takes effect, restate the prior period under the new definition, and keep the superseded edition available. A metric we quietly redefine is a metric you should stop trusting.

Section 02

What “Triage” Means Here

Triage is the interval and the judgement between a report arriving and a human case handler receiving a structured, prioritised case. It is not investigation, and it is not resolution.

SCOPE BOUNDARIES

In scope	Out of scope
Transcription of the spoken report	Investigation time and outcome
Translation into the handler’s working language	Time a handler spends reading the case
Category classification	Substantiation rate — whether an allegation proved true
Severity scoring	Disciplinary or remedial action
Routing recommendation and policy citation	Reporter satisfaction

The exclusion of substantiation is deliberate and worth stating plainly: **this benchmark measures whether the system read the report correctly, not whether the report was true.** Truth is the investigation’s job. Conflating the two is the most common way triage metrics are inflated.

Section 03

Metric 1 — Triage Latency

Definition

Elapsed wall-clock time from $\mathbf{T_0}$, the moment the reporter completes submission, to $\mathbf{T_1}$, the moment a completed case — transcript, translation where applicable, category, severity, citations and routing recommendation — is available to the assigned handler.

Reported as

- Median (p50)
- 90th, 95th and 99th percentile
- Maximum observed in the period
- Count of cases exceeding 60 seconds, with a stated cause for each

> **Why percentiles and not an average**
>
> A mean hides the cases that matter. A system that triages most reports in two seconds and occasionally stalls for six minutes has an excellent average and a real problem. We publish p95 and p99 because the tail is where a delayed high-severity report actually lives, and we publish the maximum because a single catastrophic outlier should not be able to hide inside a percentile.

Exclusions, stated in advance

- Reports abandoned before submission completes — there is no T_0.
- Cases where a customer's own configuration inserts a deliberate hold; the hold duration is excluded and the count of such cases is disclosed.
- Periods of declared platform incident are *not* excluded. Removing outages from a latency metric is the clearest form of self-flattery available to a vendor, and we will not do it.

Section 04

Metric 2 — AI/Human Classification Agreement

Definition

The rate at which the model's category assignment matches the category a qualified human reviewer independently assigns to the same report, over a defined taxonomy, where the reviewer works blind to the model's output.

Every clause in that sentence is load-bearing. **Independently** and **blind** exclude the common shortcut of asking a handler whether they agree with a label already on their screen, which measures deference rather than agreement.

Reported as

- **Raw agreement** — percentage of exact category matches.
- **Cohen's κ** — agreement corrected for what chance alone would produce. On an imbalanced taxonomy, raw agreement flatters; κ does not. We report both, and if the two diverge we explain why in that edition.

- **Per-category agreement** — because an aggregate figure can conceal a category the model handles badly. If bribery and corruption is the weakest category, that is precisely the fact a compliance officer needs.
- **Confusion pairs** — the category pairs most often mistaken for one another.

Inter-rater reliability of the humans themselves

Where two or more reviewers label the same report, we report agreement *between the humans* as well. This is the honest denominator: if two experienced compliance professionals agree with each other 82% of the time on a taxonomy, then a model agreeing with a single reviewer 84% of the time is performing at roughly human level, not at 84% of perfection. Publishing human-to-human agreement is the difference between a benchmark and an advertisement.

Section 05

Metric 3 — Severity Calibration

Definition

The relationship between the severity the model assigns and the severity the human reviewer assigns, on the same ordinal scale, measured for both magnitude and direction of error.

Reported as

- **Exact-match rate** and **within-one-level rate**.
- **Directional split** — the proportion of mismatches where the model scored *above* the human versus *below*.
- **Severe under-scoring events** — a count, published individually, of cases where the model scored two or more levels below the human. Every such case is described in the edition.

> **The asymmetry that matters**
>
> These two errors are not equivalent and must never be netted off. Over-scoring wastes a reviewer's time. Under-scoring can leave a serious allegation sitting in a low-priority queue. A system whose errors lean toward caution is behaving correctly; one with a symmetric error profile is mis-tuned for this domain. We report the direction separately for exactly this reason, and we never publish a single "calibration accuracy" figure that hides it.

Section 06

Metric 4 — Override Rate

Definition

The proportion of cases in which the handler changes the model's category, severity or routing before the case proceeds, with the reason recorded from a fixed list plus free text.

Reported as

- Override rate overall, and separately for category, severity and routing.
- Distribution of recorded reasons.
- Rate by handler tenure band, which detects deference forming over time.

This metric is not a measure of model quality, and reading it as one inverts its meaning. **An override rate approaching zero is a warning, not an achievement.** It usually means handlers have stopped genuinely reviewing — automation deference — and the human oversight the ECCP asks about has become a formality. A healthy programme shows a persistent, non-trivial override rate and a reason distribution that makes sense.

Section 07

The Human Baseline

Agreement figures are meaningless without knowing how humans perform at the same task. The baseline is established as follows, and re-established annually.

1. A stratified sample of reports is drawn across the category taxonomy and severity range.
2. Each is independently labelled by at least two qualified reviewers, blind to each other and to any model output.
3. Human-to-human agreement, both raw and κ, is computed and published. This is the baseline.
4. Model agreement is then reported *against that baseline*, never as a bare percentage.
5. Where reviewers disagree, an adjudicated label is produced by a third reviewer and the disagreement is retained in the record rather than smoothed away.

Two consequences follow, and both are stated here so no future edition can quietly avoid them. First, the model can be reported as performing *at or near human level* but not as "more accurate than humans" on a task whose ground truth is itself human judgement. Second, if human-to-human agreement is low in a category, that is a finding about the taxonomy — the categories are ambiguous — and it will be reported as such rather than blamed on the model.

> **For your ECCP evidence file**
>
> A customer can cite this section, plus the current measured edition, as the documented answer to "What baseline of human decision-making is used to assess AI?" It does not discharge the obligation to govern the technology in your own programme, but it removes the excuse that the vendor never told you.

Section 08

Sampling, Exclusions And Privacy

The benchmark is computed over real report traffic. That creates a constraint no methodology can wish away: **the underlying data is other people's misconduct reports.**

Privacy constraints on this publication

- No report content, quotation, paraphrase or identifiable detail is published in any edition.
- No customer is named or made identifiable, including by combining volume, sector and region.
- Cohorts below a minimum size are suppressed rather than published, and the suppression is disclosed.
- Human labelling for the baseline is performed under the same access controls, confidentiality obligations and audit trail as live case handling.
- Where a customer's agreement excludes its data from aggregate measurement, it is excluded.

Disclosure of the sample

Each measured edition states the number of reports in the latency population, the number in the labelled agreement sample, the category taxonomy version, the languages represented, and the number of distinct customer environments contributing — without identifying them. A sample too small to support a claim will be published with that stated, rather than aggregated across periods until it looks large enough.

Section 09

Publication Protocol

WHAT WE COMMIT TO

Commitment	Detail
Cadence	Monthly, covering the preceding calendar month, published within 15 days of month end.
First measured edition	Published once the platform has accumulated a reporting population large enough to support the sample disclosures in section 8. Until then this methodology edition stands alone, and we will not publish a figure we cannot support.
Definition changes	Disclosed in the edition where they take effect, with the prior period restated on the new definition and the old edition kept available.
Corrections	Issued as a numbered revision with the error described. Editions are never silently replaced.
Bad results	Published. A metric that only appears when it is favourable is not a benchmark. If a month is poor, the edition says so and states what changed.
Archive	Every edition remains downloadable at xposer.ai/resources.

> **Why the first edition is a methodology and not a number**
>
> We would rather publish a method we can be held to than a figure we cannot yet substantiate. The alternative — quoting an impressive number now and defining it later — is exactly the practice section 1 criticises, and we would have no standing to criticise it if we did it ourselves.

Section 10

How To Audit These Numbers

Questions we will answer, from customers and from their auditors.

- **Sample composition** — the size, stratification and language distribution of the labelled sample for a given edition.
- **Taxonomy** — the full category list and definitions in force for that edition.
- **Reviewer qualification** — the experience criteria applied to human labellers.
- **Your own environment** — latency and override figures for your tenant alone, so you can compare your experience against the published aggregate. If they diverge, we would want to know too.
- **Adjudication records** — how human disagreements were resolved, in aggregate form.

What we will not provide is report content, in any form, to anyone other than the customer that owns it. No audit of a benchmark justifies exposing a whistleblower's account.

Section 11

References

1. **Evaluation of Corporate Compliance Programs**, U.S. Department of Justice, Criminal Division, updated September 2024 — source of the human baseline question quoted in section 1.
2. **Cohen, J.** (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* 20(1), 37–46 — the κ statistic used in section 4.
3. **Landis, J. R. & Koch, G. G.** (1977), "The Measurement of Observer Agreement for Categorical Data," *Biometrics* 33(1), 159–174 — the interpretation bands conventionally applied to κ.
4. **XPOSER.AI, Anonymity Whitepaper** — the architecture within which this measurement occurs, including the privacy constraints referenced in section 8.
5. **XPOSER.AI, AI in Speak-Up Programs under the DOJ ECCP** — how these metrics fit an ECCP evidence file.

XPOSER.AI · Autonomous, voice-first AI whistleblowing. From incident to audit-ready insight in under 60 seconds — 24/7, in 100+ languages.

Methodology questions, criticism of these definitions, or a request for your own tenant figures: **info@xposer.ai**. We would genuinely rather be corrected than quoted.