Kestrel

METHODOLOGY WHITE PAPER

► INDEPENDENT METHODOLOGY · KESTREL · VERSION 4.0

What companies say. What creates value.

A conservative, rule-governed classifier that separates genuine shared value from the language of it.

VERSION	COMPANION
4.0	Strix Methodology 1.2
AUDIENCE	STATUS
Analysts, researchers, and reviewers using Kestrel	Living document — describes methodology, not a data snapshot

Contents

01 — SECTION

Executive summary

Kestrel is the qualitative engine of the platform. Where its companion, **Strix**, reads the *numbers* a mining company discloses, Kestrel reads the *words* — the sustainability, social-performance and governance narrative — and turns thousands of pages of prose into structured, comparable data.

It is built to answer one deceptively hard question: **when a company says it is creating value for society, is that claim genuine and evidenced, or is it public relations?** Sustainability reports are written to persuade. They are rich in the *language* of shared value — community, partnership, contribution, impact — and far poorer in the *substance* of it: a stated business benefit, a stated social or environmental benefit, an explicit causal link between the two, and measurable evidence that the outcome actually occurred. Kestrel's entire purpose is to separate the language from the substance, conservatively and consistently, at a scale no human reading team could match.

The design rests on a small number of commitments that are the credibility model:

1. **The model extracts; rules decide.** A language model reads each passage and extracts evidence. It does *not* assign the verdict. Deterministic, human-authored rules — versioned and applied identically to every passage — turn that evidence into a label. **No model is ever asked to decide what qualifies as shared value.**
2. **Conservative by design.** Kestrel is built to err toward false negatives, never false positives. A passage is rated as genuine shared value only when every defining criterion is *simultaneously and explicitly* present; any ambiguity resolves downward. A "positive" from Kestrel is a high bar that the disclosure cleared, not a benefit of the doubt.
3. **Only what is on the page.** Kestrel judges the text in front of it. Where a benefit or a link is merely implied, it is recorded as implied — never inferred into a fact the report did not state.

4. **Grounded in an external, published definition.** The classification scheme is not invented in-house. It operationalises Porter and Kramer's *Creating Shared Value* framework and is anchored to a hand-built codebook validated against an independent, human-coded gold standard developed in academic research.
5. **Outputs are measurements, not ground truth.** Kestrel's labels are a careful, repeatable instrument reading — reported with their uncertainty, validated against humans, and never dressed up as objective fact.

This document explains how Kestrel produces its data and how it should be read. Like its Strix companion, it is a *methodology* paper: it describes the framework, the decision architecture, and the validation, but not the internal classification rules, prompts, or model configuration that would let someone reconstruct the instrument.

---

02 — SECTION

The problem Kestrel solves

Anyone who has read a stack of mining sustainability reports knows the difficulty: the documents are long, professionally written, and engineered to leave a positive impression. Three failure modes make naïve reading — and naïve text analytics — unreliable:

- **The language of value is cheap; the substance is rare.** Almost every report talks about benefiting communities and the environment. Very few passages actually state a concrete business benefit, tie it to a concrete social or environmental benefit, and back the link with a measured outcome. A keyword or sentiment model counts the language and is fooled by it.
- **Philanthropy and risk-management masquerade as shared value.** One-off giving with no business return, internal efficiency framed as social good, and remediation of past harm dressed up as community investment all *read* like shared value. The framework Kestrel applies is specifically designed to exclude them — but only a disciplined, criteria-based reading can tell them apart at scale.

- **Generosity creeps in.** Human readers — and ungoverned models — routinely give companies the benefit of the doubt, awarding credit for outcomes the page only gestures at. That generosity is exactly what makes most "ESG sentiment" untrustworthy.

A naïve pipeline scores tone and produces a flattering number. Kestrel does the opposite: it holds every passage to an explicit, externally-defined definition of shared value and reports how few clear it — because that gap is the finding.

---

03 — SECTION

Design principles

These principles are enforced by the architecture, not left to the model's discretion. They are what make Kestrel's output defensible.

3.1 A two-layer instrument: the model extracts, rules decide

This is the central design decision. Kestrel separates *reading* from *judging*:

- **The model's job is evidence extraction.** For each passage it identifies the defining elements of shared value, the strength of the causal link claimed, and the supporting text — and reports its own confidence.
- **The rules' job is adjudication.** A deterministic, human-authored rule set converts that extracted evidence into a label. The same rules apply to every passage, every time.

The consequence is that the verdict is *auditable and reproducible*: given the same extracted evidence, the label is fixed. No passage is rated positive on a model's overall impression. This is what "rule-governed" means, and it is why the final label is more stable and more conservative than any single model judgement behind it.

3.2 Conservative by design

Kestrel is deliberately biased toward false negatives. "Positive" requires all defining criteria to be met at once and explicitly; when the evidence is partial or the language is aspirational, the passage resolves down to *ambiguous* or *negative*. In comparisons

against human coders, Kestrel is consistently the **most conservative rater** — it awards "positive" least often, buying high precision at the cost of recall. For an analyst, this is the right trade: a Kestrel positive is trustworthy, and the rate of positives is, if anything, a floor.

3.3 Text-only evidence

Kestrel scores what the disclosure says, not what it might mean. Where a business benefit or a causal link is only implied, it is graded as implied rather than promoted to a fact. When the evidence required to make a judgement is simply absent, the instrument declines to fill the gap — it records the absence rather than inferring a benefit. This is the discipline that keeps Kestrel from rewarding well-written vagueness.

3.4 Grounded in an external definition

The scheme is built on **Porter and Kramer's Creating Shared Value** (Harvard Business Review, 2011) — a recognised, externally-authored framework — and operationalised through a codebook refined against an independent, human-coded gold standard developed in doctoral research, itself cross-referenced to the ESRS disclosure standards. Kestrel is not scoring against a private, ad-hoc rubric; it is applying an established definition of what shared value is and, importantly, what it is not.

3.5 Graded, not binary

A simple positive/negative call would throw away most of what matters. Kestrel records several structured judgements on every passage (§5), so an analyst can see not just *whether* a passage clears the bar but *how* it falls short — which ingredient is missing, how strong the evidence is, which lever of shared value is in play. The nuance is the product.

3.6 Outputs are measurements, not ground truth

Kestrel's label is the output of a defined instrument, not an objective fact about the world. The methodology treats it that way: every reliability claim is reported with its uncertainty, validated against human coders, and bounded by honest limitations (§8, §10). A Kestrel reading is repeatable and defensible — not infallible.

04 — SECTION

What Kestrel measures: the Creating Shared Value test

Kestrel's headline judgement on each passage is whether it expresses **genuine, evidenced shared value**. Operationally, that means a passage must satisfy four criteria *simultaneously and explicitly*:

01 **Business benefit** A concrete benefit to the company.	02 **Social or environmental benefit** A concrete benefit beyond the company — to communities or the environment.
03 **Explicit causal linkage** The disclosure ties the two together, not merely side by side.	04 **Measurable outcome evidence** Evidence the outcome occurred — graded from assertion to verified data.

All four required — *simultaneously and explicitly*. Miss one, and the passage cannot be positive.

A passage that meets all four is **positive**. One that shows the shared-value *shape* (both benefits) but lacks an explicit link or measurable evidence is **ambiguous** — the aspiration zone. One that shows no genuine shared value is **negative**.

Why so few passages clear the bar. Most disclosure fails this test early: the large majority of passages describe no two-sided benefit at all (one-sided giving, internal efficiency, or generic narrative). Of the passages that *do* have the shared-value shape, the binding constraint is almost always **measurable outcome evidence** — companies describe *what they did* far more readily than *what changed*. The hardest, last barrier is the step from a described mechanism to a quantified outcome. This pattern — common shape, rare substance, evidence as the bottleneck — is one of the most consistent things Kestrel surfaces, and it is visible only because the criteria are applied uniformly to every passage.

05 — SECTION

The per-passage schema

Kestrel records several structured judgements on every passage. Together they let an analyst move from a single verdict to a full profile of how a company frames value.

- **Primary label** — *Positive* (genuine, evidenced shared value), *Ambiguous* (shared value claimed but not clearly evidenced), or *Negative* (no measurable shared value in the passage).
- **Initiative type** — how far an activity actually goes: *Transactional* (one-off giving or compliance), *Transitional* (improving existing operations), *Transformational* (reshaping the wider system), or *No initiative*.
- **CSV pillar** — which lever of shared value is in play, following Porter and Kramer's three: *Products & markets* (serving unmet needs), *Value-chain productivity* (greener, fairer operations), or *Local cluster development* (strengthening the surrounding economy).
- **Linkage strength** — how clearly the passage ties activity to outcome: *Explicit*, *Implied*, or *None*.
- **Evidence strength** — how hard the supporting evidence is, graded in tiers from a bare assertion up to independently verified data.

Alongside these, each passage carries the model's **confidence** and the supporting text it relied on. These roll up into company and sector profiles: the share of genuine shared value, the mix of initiative types, the dominant CSV pillars, and the strength of evidence on offer.

> **The language of value is cheap.**
> **The substance is rare.**

06 — SECTION

How the data is produced

Kestrel runs a staged pipeline. The stages below describe *what each step does*; the internal rules, prompts, and model configuration are intentionally out of scope.

STEP 01	STEP 02	STEP 03	STEP 04
Ingest	**Extract evidence**	**Adjudicate by rule**	**Aggregate**
Each report segmented into discrete passages, classified one at a time.	A frontier model reads the passage and reports the evidence — not the label.	A versioned, deterministic rule set turns that evidence into the label.	Passage labels roll up into company, sector, and time-series profiles.

1. **Ingest.** Each report is read and segmented into discrete passages of roughly 100–220 words. Each passage is classified independently — one passage per request — so no judgement is contaminated by its neighbours and every verdict is traceable to a specific span of text.
2. **Extract evidence.** A frontier large language model reads the passage and extracts the defining elements of shared value, the linkage claimed, the evidence strength, and the supporting quote — and reports its confidence. It does not assign the label.
3. **Adjudicate by rule.** A deterministic, versioned rule set turns the extracted evidence into the labels in §5. Hard disqualifiers (for example: a benefit that accrues only internally, one-sided giving with no business return, or the absence of any causal link) resolve a passage to *negative* regardless of how it is written. "Positive" is reserved for passages that pass every gate and clear the evidence threshold.
4. **Aggregate.** Passage-level labels roll up into company, sector, and time-series profiles that an analyst can compare like for like.

The decisive property of this design is that the **label is a near-deterministic function of the extracted evidence**, not a free-form model opinion. That is what makes Kestrel's output consistent across runs and defensible on audit.

07 — SECTION

The decision layer: rule-disciplined, not vibes

It is worth being explicit about why the two-layer design matters, because it is the difference between an instrument and a guess.

- **The verdict is a rule, not an impression.** Because human-authored gates and an evidence threshold do the adjudication, the same extracted evidence always yields the same label. The classifier behaves as a disciplined decision procedure, not a mood reading of the text.
- **The gates correctly exclude the look-alikes.** The framework is designed to reject the things that resemble shared value but are not it — philanthropy with no business return, internal efficiency framed as social benefit, finance-only narrative, and risk-management cleanup. Kestrel's gates enforce exactly those exclusions, uniformly.
- **The final label is more stable than its parts.** The individual sub-judgements a model makes on a single reading are noisier than the verdict built from them: the deterministic gate compresses several uncertain sub-fields into a robust call. In repeated runs, the headline label is markedly more stable than the underlying fields — a direct benefit of letting rules, not the model, make the final decision.
- **Uncertainty is self-announced.** When Kestrel is unsure, it says so: the passages whose labels move between runs are overwhelmingly the ones the model itself flagged at lower confidence, while the confident calls are stable. Confidence functions as a discipline mirror — high confidence means "this is plainly not shared value, and I can say so cleanly," and uncertainty is reserved for the genuinely contestable, shared-value-shaped passages.

08 — SECTION

Validation and reliability

Kestrel's credibility does not rest on assertion; it rests on a validation programme designed to the standard of empirical research. Because reliability figures move with every re-run and every expansion of the held-out set, this section describes the *methods* and what they establish, and leaves the current numbers to the live platform.

The human–human ceiling. The first and most important honesty: labelling shared value is a genuinely subjective task, and two trained human coders do not perfectly agree with each other. Kestrel measures that human–human agreement explicitly and treats it as the ceiling. **No classifier can be expected to exceed how well expert humans agree among themselves**, so "accuracy against humans" is interpreted against that ceiling, not against an imaginary perfect standard.

Agreement with humans. On a held-out report coded independently by human experts, Kestrel's agreement sits *inside the human–human band*. On the headline question — *is this shared value at all?* — Kestrel agrees with the human coders about as well as the coders agree with each other, and it does so as the most conservative rater in the comparison. We are candid that held-out human samples are small and that positive-class estimates from them carry wide intervals: the honest reading is that Kestrel is within the human noise band on the reports tested, not that any single agreement number is a corpus-wide guarantee.

The kappa paradox, stated plainly. Genuine shared value is rare, and at low prevalence the most familiar agreement statistic (Cohen's κ) is mathematically depressed and unstable — high raw agreement can sit alongside a low κ purely as an artefact of rarity. Rather than cherry-pick, Kestrel reports a *family* of agreement coefficients (including prevalence-robust ones) together with raw agreement and the full confusion matrices, and interprets them jointly. No single coefficient is presented as the clean truth.

Run-to-run reliability. Because the instrument uses a probabilistic model, Kestrel is re-run multiple times on the same report and the runs are treated as independent raters. Inter-run agreement is measured with Fleiss' κ and reaches *substantial* agreement on the

headline label, with the large majority of passages receiving an identical label across all runs — and, as noted, the rare disagreements are concentrated in passages the model itself flagged as uncertain. On the same report and field, Kestrel's run-to-run self-consistency sits above the inter-human ceiling.

Gold-standard development check. Kestrel is also scored against a hand-coded gold standard. We report this as a *development-set sanity check*, not a headline accuracy, and we are explicit about why: earlier versions that scored near-perfectly did so in-sample, on the very passages their codebook was tuned on — a measure of self-consistency, not accuracy. Version 4 deliberately *tightened* the bar for the ambiguous category; the resulting error profile is one-directional (formerly-ambiguous passages now resolving to negative) and reflects a chosen precision-for-recall trade, not degraded quality. The defensible cross-version comparison is the held-out human validation, on which v4 is a substantial improvement over earlier versions.

Inference discipline. When Kestrel's outputs are analysed, the statistics respect the structure of the data: passages are nested within company-years within companies, so they are *not* treated as independent observations. Comparative claims use company-level units and cluster-robust methods; passage-pooled tests that would overstate certainty are reported as descriptive only. Analyses are pre-specified and have been subjected to adversarial review. This is the same conservative instinct that governs the classifier itself, applied to the evidence about it.

---

09 — SECTION

What the lens reveals

A methodology is best judged by whether it surfaces something true and non-obvious. Applied across the sector, Kestrel consistently reveals patterns that a tone-based reading would miss. These are durable, qualitative findings; the figures behind them are maintained live in the platform.

- **Substance is rare, and the shape is common.** A large share of disclosure has the *shape* of shared value; only a small fraction carries the *substance*. The gap between the two is the central story of mining ESG narrative.
- **Companies describe activities, not outcomes.** The sector-wide bottleneck is measurable outcome evidence. Reports readily describe what was built or funded; they far less often quantify what changed as a result.
- **More talk is not more substance.** Where the language of shared value has grown over time, the substance behind it has not measurably moved — the trend, where there is one, is in the framing.
- **Remediation is often risk-management in a shared-value costume.** For firms with significant environmental exposure, a large slice of what reads as community or environmental value is cleanup of past harm — correctly distinguished from genuine, forward-looking shared value.
- **Membership labels predict little.** Belonging to an industry sustainability body is not, by itself, associated with more genuine shared-value disclosure. Kestrel lets an analyst test such claims rather than assume them.

The point is not any single statistic — it is that Kestrel makes these questions *measurable*, consistently and at scale.

---

10 — SECTION

Limitations and honest disclosures

Trust is built as much by what an instrument refuses to claim as by what it delivers.

- **Labels are classifier-defined, not ground truth.** "Positive" means *Kestrel, applying the version-4 codebook, judged the disclosure to meet every criterion*. It is a careful measurement, not an objective fact, and systematic under-labelling would deflate every rate. Because Kestrel is the most conservative rater tested, its positive rate should be read as a floor.

- **Reliability is validated on limited human-coded data.** Held-out, human-coded samples are necessarily small, and positive cases within them are rare, so agreement estimates carry wide intervals. The defensible claim is that Kestrel sits within the human–human band on the reports tested — not that any single figure generalises perfectly to every company and year.
- **Coverage is uneven and observational.** The corpus is a cross-sectional snapshot of what companies chose to publish; reporting depth and format vary by company and year, and recent years are thinner. Findings are associations within disclosure, not causal claims about real-world performance.
- **Pooled rates are volume-weighted.** A few large reporters dominate raw passage counts, so pooled rates lean toward them. Company-level views are the fair basis for cross-company comparison, and the platform foregrounds them.
- **The ambiguous category is the hardest to pin down.** The boundary between adjacent initiative types and between ambiguous and negative is the main source of disagreement — among humans and between humans and Kestrel alike — and is the primary target of ongoing refinement.
- **It reads disclosure, not reality.** Kestrel measures how a company *reports*, not how it *performs*. A thin shared-value profile may reflect modest reporting as much as modest practice; the two should not be conflated.

---

11 — SECTION

How an analyst should read Kestrel data

1. **Read "positive" as a high bar cleared, not a score.** A Kestrel positive means the disclosure met every criterion explicitly. Treat it as trustworthy and rare by design — and treat the *absence* of positives as a meaningful signal about the reporting, not a model failure.

2. **Use the ambiguous tier diagnostically.** Ambiguous passages are the aspiration zone — shared-value language without the evidence to back it. The mix of ambiguous to positive tells you how much of a company's narrative is intent versus demonstrated outcome.
3. **Look at which ingredient is missing.** The structured fields show *why* a passage fell short — usually missing measurable evidence. That is more informative than the headline label alone.
4. **Compare at the company level.** Pooled rates are volume-weighted toward the largest reporters; use the company-level views for fair comparison.
5. **Separate talk from substance over time.** Rising shared-value language does not imply rising substance. Read the evidence-strength and positive trends, not just engagement.
6. **Respect the uncertainty.** Where Kestrel flags lower confidence, the label is genuinely contestable. Treat those passages as candidates for human review, not settled fact.

A — APPENDIX

Per-passage data dictionary

Selected fields.

FIELD	MEANING
label_primary	Headline verdict: positive / ambiguous / negative.
initiative_type	Depth of the activity: no_initiative / transactional / transitional / transformational.
csv_pillar	Lever of shared value: products & markets / value-chain productivity / local cluster development / none.
linkage_strength	How clearly activity is tied to outcome: explicit / implied / none.
evidence_strength	Hardness of supporting evidence, graded in tiers (bare assertion → independently verified).
confidence	The model's self-reported confidence in its evidence extraction.
supporting quote	The verbatim text the judgement was drawn from, retained for audit.

Adjudication of these fields into the primary label is performed by the versioned rule layer, not by the model.

B — APPENDIX

Glossary

Creating Shared Value (CSV)
Porter and Kramer's framework for business activity that produces a genuine company benefit *and* a genuine social or environmental benefit, causally linked. The standard Kestrel applies.

The four criteria
Business benefit; social/environmental benefit; explicit causal linkage; measurable outcome evidence. All four are required for a positive.

The human–human ceiling
The level of agreement two expert coders reach with each other; the realistic upper bound on classifier–human agreement.

The kappa paradox
The tendency of Cohen's κ to read low when the positive class is rare, even when raw agreement is high; the reason Kestrel reports several agreement coefficients together.

Disqualifier gate
A deterministic rule that resolves a passage to negative when a defining criterion is absent (e.g. no causal link, internal-only benefit, finance-only framing).

Two-layer architecture
The separation of model-based evidence *extraction* from rule-based *adjudication*; no model decides what qualifies as shared value.

C — APPENDIX

References

- Porter, M. E. & Kramer, M. R. (2011). "Creating Shared Value." *Harvard Business Review*, 89(1), 62–77.
- Fleiss, J. L. (1971). "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, 76(5), 378–382. doi:10.1037/h0031619.
- Landis, J. R. & Koch, G. G. (1977). "The measurement of observer agreement for categorical data." *Biometrics*, 33(1), 159–174.

Kestrel Methodology v4.0 · companion to Strix Methodology 1.2. This document describes methodology, which is stable; current classification figures, agreement scores, and corpus coverage are maintained live in Kestrel. Kestrel's labels are a validated instrument reading of disclosed text — a measurement, not ground truth.