

Before You Trust the Count: Auditing Multilingual Sustainable Development Goal News

Sergey Volkov^{1,2}

¹Shenzhen Loop Area Institute

²School of Management and Economics,
The Chinese University of Hong Kong, Shenzhen
sergeyvolkov@link.cuhk.edu.cn

Abstract

News monitoring of the Sustainable Development Goals (SDGs) can be distorted by a missing archive interval, a broad lexical rule, or duplicated evidence. We study how to audit such conclusions when target-domain relevance labels are unavailable. Our case comprises 16,060,006 URL-unique candidates retrieved from English, Spanish, Russian, and Chinese GDELT Article List (GAL) titles observed in 2024–2025. We replace whole-day gap accounting with a minute-interval ledger, correct public-label processing to distinguish accepted, rejected, and unassessed text–goal pairs, and compare weak signals on source-grouped public benchmarks. We find that weak fusion does not satisfy our predefined retention rule. We therefore use a claim-level audit with three outcomes: supported under tested assumptions, reversed by a specific measurement choice, or blocked by missing evidence. Paired translated headlines reveal disagreement between language-specific instruments. Controlled engineering and semantic streams let us measure false releases, retained correct claims, and release rate against separate references. The audit lowers false-release rates on semantic tracks but abstains too often on the engineering track. The released artifact makes these tradeoffs reproducible. Agreement across specifications shows robustness to the tested checks; it does not prove semantic correctness or accuracy on unlabeled news titles.

1 Introduction

The Sustainable Development Goals (SDGs) organize policy and communication across connected development issues (United Nations, 2015). News monitoring can help analysts find relevant stories and examine changes in observed attention. Yet a fluent summary or a narrow confidence inter-

val can conceal a collection failure, an ambiguous term, or several signals repeating the same evidence. A useful monitoring system must connect each conclusion to the observations and measurement choices that support it.

We study this problem without target-corpus human relevance labels. Our question is deliberately narrower than whether an SDG classifier is accurate: *Does a specified monitoring conclusion survive a declared family of observation and measurement tests?* For example, does the measured fraction associated with SDG 9 exceed that associated with SDG 7 within a fixed observed candidate population? A comparison between languages must additionally confront the possibility that different instruments produce different outputs on the same content.

We make three contributions. First, we provide corrected observation and label accounting: a complete archive-inventory ledger with partial-day absences, and separate supported-label protocols for the OSDG Community Dataset (OSDG-CD) and SDGi Corpus. Second, we test measurement mechanisms through signal duplication, lexical attribution, counting and source influence, and paired translated headlines. Third, we provide an executable claim audit and controlled failure benchmark that reports both false releases and useful-claim retention. The benchmark explicitly separates engineering reference counts from semantic reference labels.

The contribution is an audit with visible limits. Several methods can agree and still be wrong. Public-label performance does not identify accuracy on GDELT Article List (GAL) headlines, and the controlled experiments do not establish natural-world failure rates. A negative fusion result is useful here: an additional probabilistic component must earn its place rather than become the

justification for a claim it cannot validate.

2 Related Work

Research on sustainability-related media attention predates the SDGs. [Barkemeyer et al. \(2013\)](#) compared issue priorities in 115 newspapers across 41 countries and related regional differences to socioeconomic development. [Czvetkó et al. \(2021\)](#) mapped 2019 GDELT news to the 17 goals using search terms drawn from the World Bank Group Topical Taxonomy, then analyzed country profiles, tone, and network relationships. These studies show how news-derived profiles support substantive comparisons. Our audit examines whether a specified comparison survives changes in archive availability, lexical rules, and duplicate handling before it receives that interpretation.

Existing SDG resources support different evaluation questions. The SDGi Corpus provides multilingual text classification data with labels inherited from source documents ([Skrynnyk et al., 2024](#)). OSDG-CD records judgments on proposed text-goal associations ([OSDG et al., 2024](#)); OSDG 2.0 describes a multilingual classification tool ([Pukelis et al., 2022](#)). We preserve these distinct label meanings in held-out tasks and compare supervised term frequency-inverse document frequency (TF-IDF) classifiers with weak signals. Performance on those resources does not establish relevance accuracy on news headlines.

Latent-class estimation combines imperfect observations while estimating their error rates ([Dawid and Skene, 1979](#)). Snorkel models labeling-function accuracies and dependencies to produce probabilistic training labels for discriminative models ([Ratner et al., 2020](#)). Our expectation-maximization (EM) diagnostic assumes conditional independence of signals given relevance and language. The duplication experiment tests a vulnerability of that assumption; it is not an evaluation of Snorkel’s dependency modeling. Fusion must outperform the simpler evidence-family baseline to remain central.

Measurement theory distinguishes a construct from its operationalization and the assumptions needed for validity ([Jacobs and Wallach, 2021](#)). Data-cascade research documents how upstream data problems propagate through deployed systems ([Sambasivan et al., 2021](#)). We turn these concerns into explicit claim decisions and evaluate false releases alongside retained correct claims un-

der controlled references. Parallel Global Voices supplies translation-linked documents ([Prokopidis et al., 2016](#)); paired-headline tests help separate language-instrument disagreement from differences between unrelated news populations.

3 Claims, Populations, and Observation

3.1 Fixed estimands

GAL provides minute-addressed gzip objects containing article metadata ([Leetaru, 2021](#)). We use archive observation time, not the article’s publication metadata, to define the half-open interval 2024-01-01 through 2026-01-01 UTC. These two calendar years delimit the case study; they do not establish long-run SDG trends. The corpus contains nonempty titles in GAL-detected English, Spanish, Russian, and Chinese. Candidate retrieval is language-conditioned, and exact URLs are deduplicated in observation order. Chinese language variants are mapped to one declared bucket.

Our full-corpus population is the fixed set of 16,060,006 exact-URL-unique titles matching at least one term in the frozen expanded instrument. This is an observed candidate population, not all news or all GDELT titles. The compact specification is evaluated on this same population; its non-matches stay in the denominator. For specification m , goal g , and candidate population P ,

$$\hat{p}_g^{(m)}(P) = \frac{1}{|P|} \sum_{i \in P} z_{ig}^{(m)}. \quad (1)$$

A candidate is an article unit; an assignment is one article-goal pair. Fractions in Equation 1 need not sum to one. When we normalize assignments across goals for a profile or Jensen-Shannon divergence (JSD), we name that distinct denominator explicitly.

The frozen questions compare goals 9 versus 7, 3 versus 13, and 16 versus 17, both pooled and within each language. A separate temporal question compares SDG 9 candidate fractions in June–July 2024 and June–July 2025. “Association” means broad topical relevance as operationalized by the named instrument. It does not mean explicit SDG discussion, a population’s priorities, or evidence of development progress.

Mechanism analyses use the unchanged deterministic 79,975-candidate hash sample. Those are explicitly sample-population findings. Removing a publisher changes the population; grouping normalized titles changes the counting unit. We report

Start	End	Minutes
06-12 18:02	06-12 20:16	134
06-14 17:33	07-02 01:56	24,983
10-07 19:32	10-07 20:46	74

Table 1: Flagged archive-absence intervals in 2025 (UTC). End times are exclusive; these are absent-object runs, not measured missing-news counts.

each as a separately named influence analysis and never combine them into one uncertainty interval for the publication-count estimand.

3.2 Inventory success and unknown stream coverage

The fail-closed collector resolves all 1,052,640 requested minute URLs. Only HTTP 200 and 404 are resolved inventory outcomes. Present objects must pass byte-count, gzip checksum, strict UTF-8, JSON-line, and row-count checks before a day is committed atomically. Unresolved requests are never converted to empty data. Partition reconciliation and repeated-object hash checks test the stored archive inventory.

An object need not exist at every minute: GAL objects cluster around processing batches. We therefore retain *all* contiguous absence runs and flag runs of at least 60 minutes as potential observation gaps. The threshold is an operational diagnostic, not a claim that shorter absences are harmless or that a 404 proves missing news. Intervals are half-open and include partially affected dates.

The long run in Table 1 spans two partly affected boundary days in addition to 17 wholly empty days. Two shorter flagged runs would also be missed by empty-day-only reporting. The ledger flags 21 dates in total. Successful inventory establishes which archive objects were available; completeness of the underlying news stream remains unidentified, even outside flagged intervals. The temporal claim intersects these absences and is therefore insufficiently supported.

4 Corrected Public-Label Evaluation

4.1 What the labels assess

OSDG-CD asks annotators to accept or reject a proposed text-goal association (OSDG et al., 2024). The original 2024.04 release retains both vote counts, agreement, and source DOI. For positive and negative counts v^+ and v^- , we require agreement $|v^+ - v^-|/(v^+ + v^-) \geq 0.6$ and a di-

rectional majority. Accepted and rejected pairs receive 1 and 0; other goals remain unknown. Ties and lower-agreement cases are retained separately. This yields 23,470 accepted pairs, 1,602 rejected pairs, and 17,953 uncertain cases. The strong class imbalance requires an always-accept comparison and specificity diagnostics.

The previously used label-only derivative discarded this distinction. We quarantine its evaluation, neighbor-transfer, and posterior outputs rather than relabel them as corrected results. Unknown goals are never automatically negative in the new OSDG evaluation.

SDGi is a separate supplied-label reproduction task. Its SDG labels derive from source documents and do not establish exhaustive human relevance judgments for each paragraph. Unlisted labels are zeros only for this benchmark’s supplied multilabel task. The official test set was sampled non-randomly, with classifier-loss-based weights (Skrynnyk et al., 2024). We report it separately from a new grouped split and record source overlap in the official split.

4.2 Splits, masks, and competing methods

We group OSDG excerpts by source DOI. For SDGi, type, country, locality, and year conservatively group related reports and translations. Identical normalized excerpts connect source groups before splitting. A seeded source-group hash assigns 60/20/20 train/validation/test partitions; the resulting test sets contain 8,610 OSDG excerpts and 1,637 SDGi excerpts. Known-label support, source groups, and every row’s split are released. These are grouped held-out source tasks, not tests of GAL accuracy.

Table 2 defines the five signals used in the public comparisons, paired headlines, and EM diagnostic. Public vocabularies fit training text; the character representation additionally includes frozen goal-term prototypes. Neighbor queries exclude their source group, and unknown OSDG pairs contribute no evidence. Gold fields never become lexical query features. French SDGi texts have no dedicated lexical list; activation diagnostics expose this limitation.

The frozen language-specific lists contain compact anchors based on localized UN goal terminology and broader topical additions. For example, English SDG 9 adds *technology* to the anchor *infrastructure*. The released JSON records the lists and source pages; it does not provide a term-by-

Signal	Input and computation	Output and missing-evidence behavior	Example or rule
Expanded	Text, language, compact anchors plus topical additions; regex matching.	1 for a goal-term match; otherwise 0, including languages without a lexical list.	English <i>technology</i> triggers SDG 9.
Compact	Same text and language; compact anchors only.	1 for an anchor match; otherwise 0. A nonmatch stays in the claim denominator.	<i>infrastructure</i> triggers SDG 9.
Explicit number	Text and multilingual expressions naming a numbered goal.	1 if the queried goal is explicitly named; otherwise 0. No abstention.	<i>SDG 7</i> triggers goal 7.
Prototype	Character 2–5-gram TF–IDF cosine similarity to each goal’s pooled multilingual terms.	1 for goals among the three highest positive similarities; otherwise 0. Raw cosine scores supply rankings.	Each goal has one pooled term prototype.
Public neighbor	Word 1–2-gram TF–IDF; up to five other-source training neighbors; similarity-weighted supported labels.	1 if the weighted positive fraction is at least 0.5; 0 below 0.5; –1 if no assessed neighbor evidence is available.	Unassessed OSDG goals contribute no votes.

Table 2: Signal definitions for one text–goal pair. A vote of 0 is an operational negative, not a verified rejection of topical relevance; –1 denotes abstention. Examples identify rules or keyword entries, not measured evaluation items. Public-label unknowns remain separate from signal votes.

term curation log or independent native-speaker validation. Matching uses word boundaries for alphabetic terms, declared suffix stems for Russian, and unbounded Chinese strings. Expanded terms broaden the retrieval specification; improved true recall is not established.

Every method uses the same supported-label mask. Feature abstentions make no positive prediction and rank last, with their activation also reported. We compare individual signals, an equal mean over four evidence families (expanded/compact combined once, explicit number, prototype vote, and public-neighbor score), and supervised TF–IDF logistic classifiers. A missing neighbor score contributes zero to this simple mean; the shared lexical family avoids counting expanded and compact twice. Thresholds maximize per-goal validation F1 over 0.1–0.9; test labels never select thresholds. Per-goal and per-language F1, average precision (AP), area under the receiver operating characteristic curve (AUROC), specificity, and balanced accuracy are released. Ranking metrics are undefined where a goal lacks both label classes.

4.3 Fusion as a tested hypothesis

The diagnostic EM model estimates binary latent labels and language-specific signal reliability with global shrinkage. The corrected likelihood includes the probability that a vote is observed or abstains. Its default initialization is $q_0 = 0.08 + 0.84r$, where r is the observed positive-vote fraction. We fit on training votes and apply frozen parameters to held-out texts.

We stress the 200-pseudo-observation shrinkage with levels 20 and 2,000, replace initialization

with constants 0.15 and 0.85, and duplicate the expanded lexical signal once. Convergence histories and all attempted configurations are retained. We stop when the largest posterior change is below 10^{-4} , with a cap of 200 updates. Our declared rule keeps EM central only if it converges and its validation macro F1 and AP both exceed the family mean. It does not meet this rule, so the claim audit does not depend on EM. The incomplete fits are not evidence about the performance of a fully converged optimum. The OSDG base run reaches 200 updates without meeting the tolerance (final maximum change 0.000423). The SDGi base run converges in 64 updates and the GAL sample diagnostic in 81. Appendix A reports convergence for every stress-test fit.

On OSDG, the base EM fit at the iteration cap has macro F1 0.785, compared with 0.946 for the family mean and 0.980 for supervised TF–IDF. Duplicating lexical evidence raises mean internal confidence from 0.989 to 1.000 over all queried pairs, including unassessed ones; this is a confidence diagnostic, not validation of those pairs. On SDGi, duplication changes EM macro F1 from 0.469 to 0.338. Initialization and shrinkage sensitivity are reported in Appendix A.

5 Measurement Mechanisms

5.1 Lexical, counting, and source influence

Full-corpus lexical fractions use the fixed denominator in Equation 1. Table 4 connects representative claims to a concrete mechanism or missing requirement. Some orderings persist but fall below the predefined margin; agreement in direction alone is not enough for release.

Method	OSDG F1	AP	BA	SDGi F1	AP	BA
Always accept	0.973	0.949	0.500	0.189	0.105	0.500
Expanded	0.764	0.980	0.796	0.337	0.188	0.663
Compact	0.342	0.959	0.616	0.360	0.223	0.638
Explicit number	0.005	0.949	0.501	0.434	0.296	0.653
Prototype	0.608	0.990	0.728	0.380	0.437	0.631
Public neighbor	0.855	0.978	0.781	0.622	0.610	0.738
Family mean	0.946	0.991	0.774	0.577	0.577	0.717
EM	0.785 [†]	0.992	0.806	0.469	0.550	0.694
Supervised TF-IDF	0.980	0.997	0.610	0.672	0.774	0.767

Table 3: Grouped held-out public tasks, macro averages over supported goals. OSDG assesses accepted/rejected proposed pairs; SDGi reproduces supplied document-derived label vectors. AP is high on imbalanced OSDG even for always-accept; balanced accuracy (BA) exposes this. No value estimates GAL accuracy. [†] The OSDG EM fit reached the iteration cap without convergence.

Claim	Expanded	Compact	Audit
Full: SDG 9 > 7	11.12	0.82	Insufficient
Full: SDG 3 > 13	14.82	7.75	Supported
Full: SDG 16 > 17	1.18	1.33	Insufficient
Sample: SDG 16 > 17	1.17	1.35	Sensitive

Table 4: Pooled signed candidate-fraction margins (percentage points). Full and sample populations are separate. The sample SDG 16/17 ordering reverses when the English term *justice* is removed; the temporal comparison is insufficient because of unknown observation coverage.

For every language-goal term in the hash sample, we record assignments it triggers and assignments for which it is the sole trigger. Leaving out a term removes an assignment only in the latter case. Removing an entire topical-addition family recovers the compact instrument for that goal. This provides an inspectable counterexample rather than an unexplained change in an aggregate score.

Exact-URL counts, recognized-tracking-parameter canonicalization, and normalized-title story proxies are separately named analyses. A normalized title does not establish that two publications describe the same story, and different titles may still be duplicates. We also measure publisher concentration and exclude each stratum’s largest domain as a restricted-population influence analysis. These checks identify dependence on counting and source composition; they do not repair unknown semantic errors.

To examine retrieval blind spots, we select archive chunks before reading titles: the first present object in each UTC hour on eight predefined dates. All 192 objects match their stored hashes and row counts. A 10,000-title hash sample from their four-language title stream is scored be-

fore retrieval. Disagreement outside the expanded candidate set is a diagnostic about instrument behavior, never a measurement of true retrieval recall. Public or prototype activation can itself be wrong.

5.2 Paired multilingual headlines

We select up to 1,000 English–Spanish, English–Russian, and English–Chinese document pairs by a fixed hash of the pair IDs in Parallel Global Voices (Prokopidis et al., 2016). Titles come from explicit title elements in the archived XML; missing or empty titles and ambiguous links are counted. Both sides pass through retrieval and goal assignment. We evaluate expanded and compact retrieval with lexical, family-mean, and diagnostic EM assignment, and repeat the comparison on one-to-one links.

We report retrieval disagreement, any-goal assignment disagreement, disagreement among pairs active on either side, and normalized assignment-profile JSD. Table 5 includes activation so an instrument cannot appear useful merely by predicting nothing. English is not truth: translations may localize headlines, and link-based pairing can be imperfect. The older paired corpus also differs from the GAL period and source stream. Compact retrieval lowers disagreement on all pairs while sharply reducing activation. It does not uniformly improve assignment agreement: on English–Russian pairs, active-union disagreement changes from 46.6% to 54.8%. Thus the compact instrument is a sensitivity specification, not a demonstrated multilingual correction.

Pair	Lexicon	n	E act.	T act.	RD	AD
en-es	expanded	1000	9.0	8.1	2.9	33.0
en-es	compact	1000	1.7	1.7	0.4	21.1
en-ru	expanded	1000	12.7	13.0	6.5	46.6
en-ru	compact	1000	2.4	3.8	2.2	54.8
en-zh	expanded	1000	11.0	7.6	10.2	72.9
en-zh	compact	1000	2.5	1.4	1.9	65.5

Table 5: Paired headline diagnostics (percentages). Activation is shown separately on the English (E) and translated (T) sides. Retrieval disagreement (RD) uses all pairs; assignment disagreement (AD) conditions on either side being retrieved. Neither language is ground truth.

6 Executable Claim Audit

Let $\mathcal{M}$ be the declared specifications for one fixed question and population. A claimed ordering has signed margin d_m under specification m . The rule requires at least 30 observed units and a margin above 0.02 for every specification. Where a trusted benchmark manifest supplies a bound on missing units, it additionally requires $(nd_m - u)/(n + u) > 0.02$, where n is the observed unit count and u bounds the missing units. Each missing unit can oppose an ordering by at most one. Natural GAL absences do not supply a missing-record count; their temporal comparisons remain blocked.

The output is one of three states:

- **Supported under the tested assumptions:** all required evidence is present and every specification preserves the claimed direction above the margin.
- **Specification-sensitive:** a concrete specification reverses or ties a direction supported by another specification.
- **Insufficient evidence:** coverage, sample size, or margin is inadequate, or the proposed direction is unsupported throughout.

The report contains the fixed population, unit, denominator, signed margins, and either a counterexample or the missing requirement. Specifications and thresholds were recorded before the new benchmark test results. This is a reproducible analysis plan, not prospective external preregistration: the prior manuscript’s aggregate directions were already visible. Agreement is not converted into a calibrated probability of correctness.

7 Controlled Failure Evaluation

7.1 Two reference tracks

The **engineering track** starts from observed GAL chunks under a fixed expanded-lexicon and exact-URL counting rule. We sample source chunks and then 120 unfiltered title units. The original instrument output is the reference. Hiding chunks, replaying a concentrated block, and introducing tracking-parameter URL variants can distort that output. This tests preservation of a specified measurement, not true SDG relevance. Four 2024 dates supply development pools; four 2025 dates supply final-test pools.

The **semantic track** constructs two 120-example periods from held-out labeled texts for each supported queried goal. OSDG uses only accepted/rejected pairs for that goal; SDGi uses its separate supplied-label vector. Positive fractions are (0.4,0.4), (0.3,0.7), or (0.7,0.3), creating unchanged and genuinely changed compositions under the benchmark label reference. A goal requires at least 30 positive and 30 negative examples in its split. Only SDG 16 meets this requirement in the OSDG final pool; SDGi supports goals 3, 7, 9, 13, and 16. Small strata may be resampled with replacement, with original row and source IDs retained. These are controlled compositions, not estimates of a natural temporal process.

The ordinary semantic pipeline proposes material increase, material decrease, or no material change using a 0.10 change threshold on supervised predictions. The audit checks the same proposed conclusion against family-mean, public-neighbor, and canonical-URL specifications. Training source groups and final stream source groups remain separated. Technical corruption is applied after stream construction; the reference labels are used only for scoring correctness.

Final tests hold out source groups, corruption severities (0.1/0.3/0.6), and a hide-plus-replay

combination. Twenty seeded replicates include unchanged streams, no corruption, and anomalies that do not change the reference conclusion. Detailed results retain the distinction between detecting a technical anomaly and preventing a consequential error.

7.2 Utility and abstention

We compare release-all, coverage-only, simple specification agreement, the complete audit, and always-abstain. Removing the coverage or specification component supplies ablations. A method’s release decision is evaluated on the *same candidate conclusions*, not on a self-selected easier truth definition:

$$\text{FRR} = \frac{\text{incorrect released claims}}{\text{all released claims}}, \quad (2)$$

$$\text{Retention} = \frac{\text{correct released claims}}{\text{all correct candidate claims}}. \quad (3)$$

We also report release rate. FRR is undefined for zero releases; always-abstain cannot receive a fabricated zero error rate. Matched-release comparisons thin eligible baseline releases, without using correctness, to the audit’s release count over 50 fixed seeds. These comparisons condition on a finite designed benchmark; they are not guarantees for unseen failures.

On the engineering track (780 claims), FRR falls from 5.8% with release-all to 0.0% with the audit. The audit retains 1.0% of correct candidate claims and releases 0.9% of all claims.

On the OSDG semantic track (780 claims), FRR falls from 35.0% with release-all to 1.5% with the audit. The audit retains 25.6% of correct candidate claims and releases 16.9% of all claims.

On the SDGi semantic track (3,900 claims), FRR falls from 26.4% with release-all to 0.0% with the audit. The audit retains 37.1% of correct candidate claims and releases 27.3% of all claims.

These measured tradeoffs do not establish correctness of GAL conclusions. The appendix reports matched-release comparisons; corruption-specific records distinguish harmless anomalies from changed conclusions. Shared specification failures and loss of correct claims remain explicit failure modes.

The engineering result is a failure of useful retention: the full audit releases only seven claims and retains less than one percent of correct candidate claims. Requiring compact semantic agreement is excessively restrictive when the refer-

ence is the fixed expanded instrument. Coverage-only retains substantially more correct engineering claims, although it permits some errors. On SDGi, specification agreement retains more correct claims than the full audit at a small observed false-release cost. Thus the complete rule is not uniformly preferable; coverage and semantic specification checks answer different questions. We retain the frozen test outcome rather than retune the rule after seeing it. Zero observed errors in a finite, dependent benchmark does not establish zero risk.

8 Reuse and Operational Meaning

The versioned artifact includes claim specifications, group-split manifests, label masks and support counts, corruption definitions, source hashes, measured claim records, regression fixtures, and commands generating every included table and figure. A small offline demonstration accepts a JSON list of claims with named populations, units, denominators, and specification margins. It returns the same three-way evidence report without the local GAL database. The full path downloads pinned public sources and reconstructs the larger experiments.

For an analyst, the artifact answers a concrete question: would a supplied monitoring conclusion survive these declared tests? It does not infer missing labels or guarantee a conclusion that passed. Full-corpus posterior scoring is unnecessary for this contribution: lexical summaries are exact for the fixed corpus, whereas the regenerated model outputs and mechanism checks use the declared sample. We do not promise a corpus-wide posterior review queue.

9 Conclusion

Monitoring conclusions need an explicit question, fixed denominator, and a record of the assumptions that can overturn them. Correct interval accounting exposes partial-day absences; correct vote processing makes assessed rejections negative and leaves unassessed OSDG pairs unknown. Fusion stress tests and paired headlines reveal instrument dependence that aggregate confidence alone can hide. The controlled benchmark evaluates the resulting audit through false releases and retained correct claims together. Its output is an inspectable robustness judgment for the stated tests, while target-domain accuracy on GAL remains unidentified.

Method	Eng. FRR	Ret.	Rel.	OSDG FRR	Ret.	Rel.	SDGi FRR	Ret.	Rel.
Release all	5.8	100.0	100.0	35.0	100.0	100.0	26.4	100.0	100.0
Coverage only	1.4	56.3	53.8	39.0	50.5	53.8	30.6	50.8	53.8
Specification agreement	0.0	3.3	3.1	5.7	45.8	31.5	0.9	73.1	54.2
Full audit	0.0	1.0	0.9	1.5	25.6	16.9	0.0	37.1	27.3
Always abstain	—	0.0	0.0	—	0.0	0.0	—	0.0	0.0
Without coverage	0.0	3.3	3.1	5.7	45.8	31.5	0.9	73.1	54.2
Without specifications	4.5	11.4	11.3	36.1	46.2	46.9	28.3	47.5	48.8

Table 6: Final controlled benchmark results (%). Each track reports false-release rate (FRR), correct-claim retention (Ret.), and release rate (Rel.). Dashes denote undefined FRR when nothing is released. Reference meanings differ by track. The coverage ablation equals specification agreement by construction.

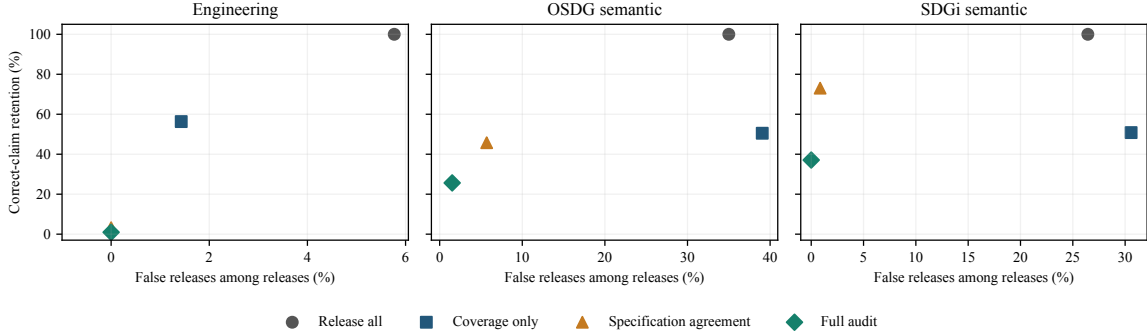


Figure 1: False-release rate and correct-claim retention on each final controlled track. Points are measured outcomes under the frozen rule. A lower false-release rate must be assessed together with retained correct conclusions. Zero-release methods have undefined false-release rate and are reported in Table 6.

Limitations

No GAL human relevance labels are available. Target precision, recall, calibration, and retrieval recall remain unidentified; public-source performance cannot fill that gap. Agreement among lexical and learned specifications can share a systematic error. The 60-minute archive-absence flag is a diagnostic threshold; source-stream completeness cannot be established even when inventory checks pass. The monitoring population excludes unobserved news, empty titles, other languages, and titles missed by the expanded lexicon.

OSDG consensus is itself an imperfect reference, with strong accepted-pair imbalance and unknown unassessed goals. SDGi labels are inherited from source documents, and its official test is deliberately non-random. Grouping by available provenance cannot exclude every unrecorded relationship. French lacks a dedicated lexical instrument. The paired news corpus is older and its translated headlines can differ in content; automated links are imperfect.

Benchmark corruptions and label compositions are controlled designs. They do not estimate the prevalence or severity of operational failures. Re-

peated streams reuse fixed held-out pools, sometimes with replacement, and should not be treated as independent new corpora. Matched-release thinning measures performance on these finite claims, not universal risk control. The specifications, 30-unit requirement, and margins are inspectable operational choices, not calibrated guarantees. Normalized-title story grouping is only a duplicate proxy, and publisher removal changes the population.

Ethical and Societal Implications

Monitoring summaries can redirect attention or encourage unsupported comparisons among communities. Here, positive impact means making a misleading release less likely while preserving useful conclusions under an explicit, testable reference. We evaluate that tradeoff rather than infer benefit from the presence of an SDG label. No deployment study measures downstream policy outcomes, so we do not claim demonstrated societal benefit.

We use public metadata and research corpora without profiling individuals. Titles can mention sensitive events or people; reuse should preserve attribution and appropriate access and retention

practices. Poor language-specific activation must remain visible. Withholding a claim reflects insufficient evidence, not a community’s importance. Redistribution follows each dataset’s license, including SDGi’s noncommercial share-alike conditions.

Acknowledgments

Generative AI tools assisted with research-code implementation, debugging, and manuscript drafting and revision. Reported measurements come from the documented computational runs. The author is responsible for the final methods, results, citations, and text.

References

- Ralf Barkemeyer, Frank Figge, and Diane Holt. 2013. Sustainability-related media coverage and socioeconomic development: A regional and North–South perspective. *Environment and Planning C: Government and Policy*, 31(4):716–740. doi:10.1068/c11176j.
- Tímea Czvetkó, Gergely Honti, Viktor Sebestyén, and János Abonyi. 2021. The intertwining of world news with Sustainable Development Goals: An effective monitoring tool. *Heliyon*, 7(2):e06174. doi:10.1016/j.heliyon.2021.e06174.
- A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28. doi:10.2307/2346806.
- Kalev Leetaru. 2021. Announcing the GDELT Article List & RSS Feed. *The GDELT Project*, December 6. <https://blog.gdelproject.org/announcing-the-gdelt-article-list-rss-feed/>
- Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385. Association for Computing Machinery. doi:10.1145/3442188.3445901.
- OSDG, UNDP IICPSD SDG AI Lab, and PPMI. 2024. OSDG Community Dataset (OSDG-CD), version 2024.04. Zenodo. <https://doi.org/10.5281/zenodo.11441197>.
- Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel Global Voices: A collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 900–905. European Language Resources Association. <https://aclanthology.org/L16-1144/>.
- Lukas Pukelis, Nuria Bautista-Puig, Gustė Statulevičiūtė, Vilius Stančiauskas, Gokhan Dikmener, and Dina Akyzbekova. 2022. OSDG 2.0: A multilingual tool for classifying text data by UN Sustainable Development Goals (SDGs). *arXiv:2211.11252*. doi:10.48550/arXiv.2211.11252.
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2–3):709–730. doi:10.1007/s00778-019-00552-1.
- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Moïs Aroyo. 2021. “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15. Association for Computing Machinery. doi:10.1145/3411764.3445518.
- Mykola Skrynnyk, Gedion Disassa, Andrey Krachkov, and Janine DeVera. 2024. SDGi Corpus: A comprehensive multilingual dataset for text classification by Sustainable Development Goals. In *Proceedings of the 2nd Symposium on NLP for Social Good*, CEUR Workshop Proceedings, volume 3764.
- United Nations. 2015. *Transforming Our World: The 2030 Agenda for Sustainable Development*. General Assembly Resolution 70/1.

A Additional Diagnostics

Table 7 reports all EM stress-test fits, including convergence status. Table 8 compares methods at matched release counts; Table 9 records supported test pairs and source overlap.

B Reproducibility Statement

All results in this revision are generated in an isolated version-2 output namespace. Prior artifacts and the invalidation ledger are preserved separately. Fixed seeds, input hashes, split membership, thresholds, model convergence, per-goal metrics, and per-claim benchmark outcomes are recorded. The released demonstration runs without credentials, GPUs, or the full local database; full reproduction uses the documented public downloads and the optional GAL recollection path. No target labels, illustrative percentages, or fabricated model outputs are presented as measurements.

Setting	OSDG F1	AP	Iter.	Conv.	SDGi F1	AP	Iter.	Conv.
Base EM	0.785	0.992	200	No	0.469	0.550	64	Yes
Prior 20	0.800	0.992	151	Yes	0.500	0.547	71	Yes
Prior 2,000	0.711	0.992	73	Yes	0.534	0.578	26	Yes
Initialization 0.15	0.801	0.992	200	No	0.506	0.559	94	Yes
Initialization 0.85	0.857	0.988	200	No	0.125	0.070	91	Yes
Duplicated expanded	0.764	0.992	18	Yes	0.338	0.529	13	Yes

Table 7: EM stress tests on grouped held-out tasks. Low/high initialization is 0.15/0.85; shrinkage varies around the 200 baseline. Iter. is the update count; Conv. indicates whether the maximum posterior change fell below 10^{-4} . Unconverged fits are diagnostic, not comparisons of fully converged optima.

Track	Method	Released	FRR (%)	Ret. (%)
Engineering	Release all	7	6.6	0.9
Engineering	Coverage only	7	2.9	0.9
Engineering	Specification agreement	7	0.0	1.0
Engineering	Full audit	7	0.0	1.0
OSDG semantic	Release all	132	35.5	16.8
OSDG semantic	Coverage only	132	39.0	15.9
OSDG semantic	Specification agreement	132	5.8	24.5
OSDG semantic	Full audit	132	1.5	25.6
SDGi semantic	Release all	1064	26.5	27.3
SDGi semantic	Coverage only	1064	30.5	25.8
SDGi semantic	Specification agreement	1064	0.9	36.8
SDGi semantic	Full audit	1064	0.0	37.1

Table 8: Matched release counts on final claims. Means use 50 seeded, outcome-blind thinnings of each eligible baseline. This is a finite-benchmark comparison, not a calibrated selective-risk guarantee.

Task	Texts	Positive	Negative	Overlap
OSDG grouped	8610	4755	281	0
SDGi grouped	1637	2919	24910	0
SDGi official	1470	2105	22885	356

Table 9: Test support by task. SDGi negatives mean unlisted supplied labels; OSDG negatives are assessed rejections. The final column counts source groups shared by official train and test, which are absent in the grouped tasks.