Interactive Explainer for Overlapping Labels in Multi-Label SDG Classification

Adrian M.P. Braşoveanu[1], Albert Weichselbraun[2,3], Lyndon J.B. Nixon[1,4], Arno Scharl[1,3]
[1]Modul University Vienna, Austria
[2]University of Applied Sciences of the Grisons, Chur, Switzerland
[3]webLyzard technology, Vienna, Austria
[4]Modul Technology, Vienna, Austria
adrian.brasoveanu@modul.ac.at

***Abstract*—Multi-label tagging of research and policy documents against the 17 UN Sustainable Development Goals (SDGs) remains an open problem: even modern Large Language Models (LLMs) paired with knowledge graph-based post-correction leave a substantial residual error. We argue that part of this gap may reflect not a model deficiency but an artefact of how the 17 SDGs were constructed, through political negotiation among 193 UN member states rather than along empirical category boundaries. We present an interactive document explainer that surfaces the resulting class entanglement through three heterogeneous lenses: lexical overlap in the document text, label-correlation structure across the corpus, and label co-occurrence among nearest research neighbours. Pinning any SDG displays its entanglement profile across all three lenses simultaneously, enabling a per-document audit of the classes the classifier could not separate. We illustrate the explainer on representative papers from the Aurora corpus, where the three lenses converge on the same conclusion: the SDGs the classifier missed and those it found correctly cluster together across all three lenses.**

***Index Terms*—Data visualization, explainable AI, human computer interaction, large language models, multi label classification, nearest neighbor methods, sustainable development**

I. Introduction

The 17 UN Sustainable Development Goals [1] are a multi-label scientific text classification benchmark adopted by major reporting initiatives. The full SDG framework is two-tiered: 17 broad *goals* (e.g. Goal 5 "Gender Equality"), each operationalised by a small number of more specific *targets* (e.g. Target 5.5 "ensure women's full and effective participation ... in political, economic, and public life"), 169 in total. Tagging research papers, policy briefs, and project reports with the goals supports a range of downstream uses: institutional reporting, funding allocation, literature review, and policy targeting [2]. The applied need is real: manual annotation does not scale, expert annotators disagree on roughly one in five papers, and the most-cited SDG mapping resources publish only the highly aggregated 17-goal labels rather than the 169-target detail.

The natural language processing response has been to push the usual levers: better backbones, careful threshold calibration, label-correlation regularisers, and recently knowledge graph-based correction of classifier outputs against the Pradhan [3] SDG interaction matrix. Each lever delivers incremental gains. None closes the residual gap to gold labels.

We argue that a substantial part of that gap may be less a classifier problem than a *class construction* problem. The 17 SDGs were finalised in 2015 after multi-year negotiations among 193 member states, with explicit attention to balancing political constituencies across the main topics [4]. They were not drawn at the joints of natural data. Climate (SDG 13) and clean energy (SDG 7) overlap because climate mitigation *is*, to a high degree, an energy transition. Gender equality (SDG 5) and reduced inequalities (SDG 10) overlap because gender disparity *is*, in development contexts, an axis of inequality. The model is not "wrong" when it confuses these classes; the classes themselves are not separable.

The visualisation contribution of this paper is an interactive tool that surfaces this entanglement. Our Explainer takes the output of a state-of-the-art SDG classification pipeline and exposes, for any given document, three independent views of which classes are tangled: i) a lexical lens (the keyword spans in the abstract that simultaneously tag two or more SDGs), ii) a corpus-correlation lens (the chosen SDG's row of an empirical 17×17 label-correlation matrix, for which a published interlinkage matrix such as Pradhan's [3] can be substituted with no other code change), and iii) a co-label lens (the rate at which the chosen SDG is co-labelled with each other SDG, computed both across the loaded research corpus and among the document's SBERT-nearest neighbours).

A single click on any SDG, regardless of location (e.g., in the abstract, in the comparison grid, or on the network), pins it across every panel and triggers the three-lens drill-in. When all three lenses converge on the same handful of neighbour SDGs, the user sees, with their own clicks, a signature consistent with the political construction of the class taxonomy.

Our intended audience is twofold. For machine learning researchers, the tool reframes the residual error in SDG classification as a class-design problem rather than a calibration problem. For sustainability domain experts, it provides a way to interrogate individual model predictions in language that connects to the substantive interlinkage literature [3], [5].

The remainder of the paper contextualises this contribution against related work (§II), describes the underlying classifier (§III), discusses the explainer itself and its three lenses (§IV), and concludes with a discussion on potential expansions (§V).

II. Related Work

a) Per-instance audit of multi-label classifiers: Generic post-hoc explanation methods such as LIME [6] and SHAP [7] attribute a single class prediction to input features. They are not designed for label *co-incidence*: when two predicted labels share the same evidence, attribution methods report two separate but largely identical heatmaps rather than the substantive observation that the labels themselves are entangled. Our explainer is in a complementary family of tools that takes label structure as a first-class object of inspection rather than a sum of independent predictions.

b) Visualisation of multi-class confusion: Confusion-matrix and class-similarity visualisations have a long history in visual analytics for classification [8], and instance-level performance-analysis tools such as ModelTracker [9] let practitioners inspect where a model errs on individual examples. Such tools surface aggregate confusion across a held-out set; they answer "which classes does the model confuse on average?". Our drill-in answers a different question: "for this specific document, which classes are entangled, by which independent evidence sources, and to what degree?". The unit of analysis is one document and one pinned class.

c) SDG interlinkage visualisation: The sustainability community has produced several interlinkage visualisations, including the IGES interactive matrix tool that exposes the 169×169 target-level matrix [10] and various network renderings of the Pradhan and ICSU goal-level scores. These tools are designed for policy analysts examining the goal taxonomy abstractly. They are not connected to any document classifier and do not surface document-conditional evidence. Our explainer borrows the visual vocabulary of signed coloured edges from this lineage but binds each interaction edge to a specific document and a specific classifier output.

d) Visual analytics for SDG and environmental communication: A separate lineage of work has developed visual analytics platforms for SDG-relevant content streams. The UNEP Live platform aggregates news, social media, and policy content and aligns them with environmental indicators in support of SDG communication; Scharl et al. [11] generalise the underlying design to lexical, geospatial, and relational context lenses. These systems target stakeholder monitoring at the corpus level. Our explainer borrows the heterogeneous-lens design philosophy but inverts the unit of analysis: a single document and a single pinned class, with lenses chosen to diagnose classifier-versus-class-construction error rather than to track public discourse.

e) Plurality and inconsistency of SDG mappings: Bibliometric literature establishes empirically that no single agreed-upon mapping from research output to the SDGs exists. Armitage et al. [12] and Purnell [13] report that independently built query- and model-based mappings retrieve largely non-overlapping publication sets for the same goals; Mutz et al. [14] find systematic disagreement across four mapping systems on over half a million papers; and Matsui et al. [15] show that prominent labelling systems differ markedly in sensitivity and specificity and carry per-goal biases. Rafols et al. [16] draw the methodological conclusion that mapping should expose a *plural* landscape rather than one authoritative assignment. This body of work supplies independent, corpus-scale evidence for the premise our explainer builds on—that SDG label boundaries are interpretation-dependent—but it operates on whole mapping systems and aggregate corpora. Our contribution is complementary: we bring the same entanglement down to a single document and a single pinned class, bound to a specific classifier output.

f) Politically-constructed taxonomies as ML problems: Critiques of the 17-goal aggregation are extensive in sustainability scholarship [4], [5], [17], but to our knowledge have not entered the multi-label NLP literature as a hypothesis about residual classifier error; we make that argument visually accessible at the per-document level.

III. Method

The explainer visualises the output of a two-stage SDG classification pipeline run on a sample of the Aurora multi-label corpus [2] (127,398 papers, a 10,000-document test split, a 500-document validation set, and 40 few-shot exemplars). The explainer itself is a single, self-contained HTML file with vanilla JavaScript and no dependencies, and it can be deployed by dropping a generated data file alongside it. It accepts any suitably-shaped JSON export from the Aurora dataset or similar multi-class SDG classification pipelines, including deep-learning SDG text classifiers such as SDG-Meter [18].

A. Two correction methods

A locally-served, NVFP4-quantised Gemma 4 26B A4B model (`Gemma-4-26B-A4B-it-NVFP4`) maps each abstract to per-target probabilities across the 169 official SDG targets. We evaluate two alternative post-correction methods on the LLM output.

a) Graph method (G-signed): Goal-level probabilities $p \in [0,1]^{17}$ are aggregated from the target-level scores and then propagated on a row-normalised, signed version of the Pradhan 17×17 indicator-correlation matrix W. The update rule is

$$p'_j = \alpha\, p_j + (1-\alpha)\sum_k W^{+}_{jk}\hat{p}_k - (1-\alpha)\sum_k |W^{-}_{jk}|\,\hat{p}_k \quad (1)$$

where $\hat{p}_k = p_k \cdot \mathbf{1}[p_k \geq \tau]$ is a confidence-gated source and α, τ are calibrated independently for each variant of W on the validation set.

b) Per-document method: A target-aggregated baseline is mixed per-goal with two document-conditional evidence vectors: SBERT-similarity-weighted neighbour labels drawn from a 5,000-paper out-of-fewshot pool, and Jaccard overlap between LLM-extracted abstract entities and per-target keyword profiles. For each of the 17 goals we sweep $\lambda_{\mathrm{nbr}}, \lambda_{\mathrm{ent}} \in [0,1]$ on a grid (subject to $\lambda_{\mathrm{nbr}} + \lambda_{\mathrm{ent}} \leq 1$) and pick the pair that maximises mean F_1 across the five validation folds; this yields one pair per goal, capturing the fact that some goals (e.g. SDG 17) benefit more from neighbour evidence than others.

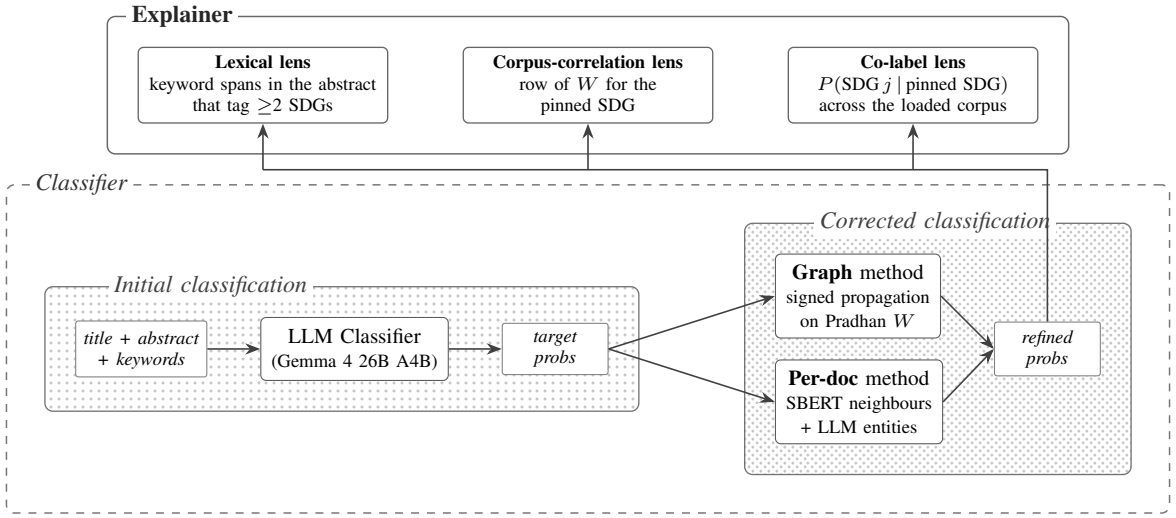


Fig. 1. The *Classifier* two-stage architecture and the *Explainer*. *Initial classification* (sparse-dot fill) produces target-level probabilities from a locally served LLM; *Corrected classification* (dense-dot fill) offers two alternative post-corrections: a graph method that propagates on a signed Pradhan matrix, and a per-document method that mixes the LLM output with SBERT-neighbour and LLM-entity evidence. The *Explainer* (upper box) consumes the refined per-goal probabilities and exposes them through three independent lenses (Section IV).

B. Data and implementation provenance

Two data scopes must not be conflated. The *classifier* is evaluated on the full Aurora silver-label corpus [2] (127,398 papers): a 10,000-document test split, a 500-document validation set used only for threshold and mixing-weight calibration, 40 few-shot exemplars held out from both, and a separate 5,000-paper out-of-few-shot pool from which SBERT neighbours are drawn. The *explainer*, by contrast, visualises only a small JSON export dropped beside the HTML—a few dozen to a few hundred documents—not the full corpus. The corpus-correlation and co-label lenses are therefore both computed over this loaded export by default (the SBERT-neighbour sub-panel additionally draws candidates from the 5,000-paper pool); their estimates inherit the small-sample noise of an export this size, which is why the per-document co-label counts are flagged as small and which a substituted published matrix such as Pradhan's removes for the correlation lens. Aurora labels are *silver* (assigned by the published Aurora target queries, not by expert adjudication), so reported numbers agree with silver labels and inherit their query biases. The per-class keyword dictionary stems from UN indicator metadata and is fixed across documents; abstract entities are extracted by the same Gemma model and matched to per-target keyword profiles by Jaccard overlap. The decision threshold defaults to 0.5; the graph weights α, τ and the per-goal mixing weights $\lambda_{\mathrm{nbr}}, \lambda_{\mathrm{ent}}$ are fit on the validation folds and then frozen.

C. Quantitative anchor

A minimal evaluation summary motivates the visualisation. On the 10,000-document test set, the goal-level LLM baseline reaches macro-F_1 = 0.522; aggregating from the target-level (169-class) outputs raises this to 0.529 (the target-aggregated ceiling). Among the graph-based corrections, no variant improves on the baseline: the goal-level Pradhan matrix reaches only 0.519, and the strongest variant (semantic clusters) reaches 0.526, within the random-matrix floor (0.525) and statistically indistinguishable from applying no correction at all. The per-document neighbour method, in contrast, reaches 0.593 (+7.1 pp over the 0.522 LLM baseline), and the combined neighbour–entity method reaches 0.589. The largest per-goal gains from neighbours are concentrated on SDG 17 ($+21.3$ pp) and SDG 8 ($+15.9$ pp); the smallest are on SDGs 7, 6, and 3, where the LLM is already strong. Read as a source ablation, this comparison isolates where the gain originates: neighbour evidence carries almost all of it, while adding entity overlap on top yields no further improvement (0.589 vs. 0.593). *The graph and per-doc corrections target different failure modes, and the explainer is the artefact that makes that difference legible per document.*

IV. The SDG Document Explainer

A. Layout

Figure 2 shows the dashboard layout, deliberately tabular: a fixed left column for the document under inspection, a fixed top-right panel for the prediction comparison, and a tabbed lower panel that toggles between two visualisations (*Charts* and *Method details*). The **document panel** renders the title, DOI, and abstract; each substring matching one of the 17 keyword sets in the Aurora dictionary is wrapped in a coloured span, and spans that match more than one SDG are merged so a single span can carry multiple affiliations. A row of Aurora silver-label chips below the abstract provides ground truth. The **prediction comparison** grid renders three rows—Aurora silver labels, LLM thresholded predictions, and the chosen correction method—with a side annotation summarising threshold crossings in natural language: which SDG was lifted by which contributing SDG, or which false positive was suppressed by which trade-off edge.

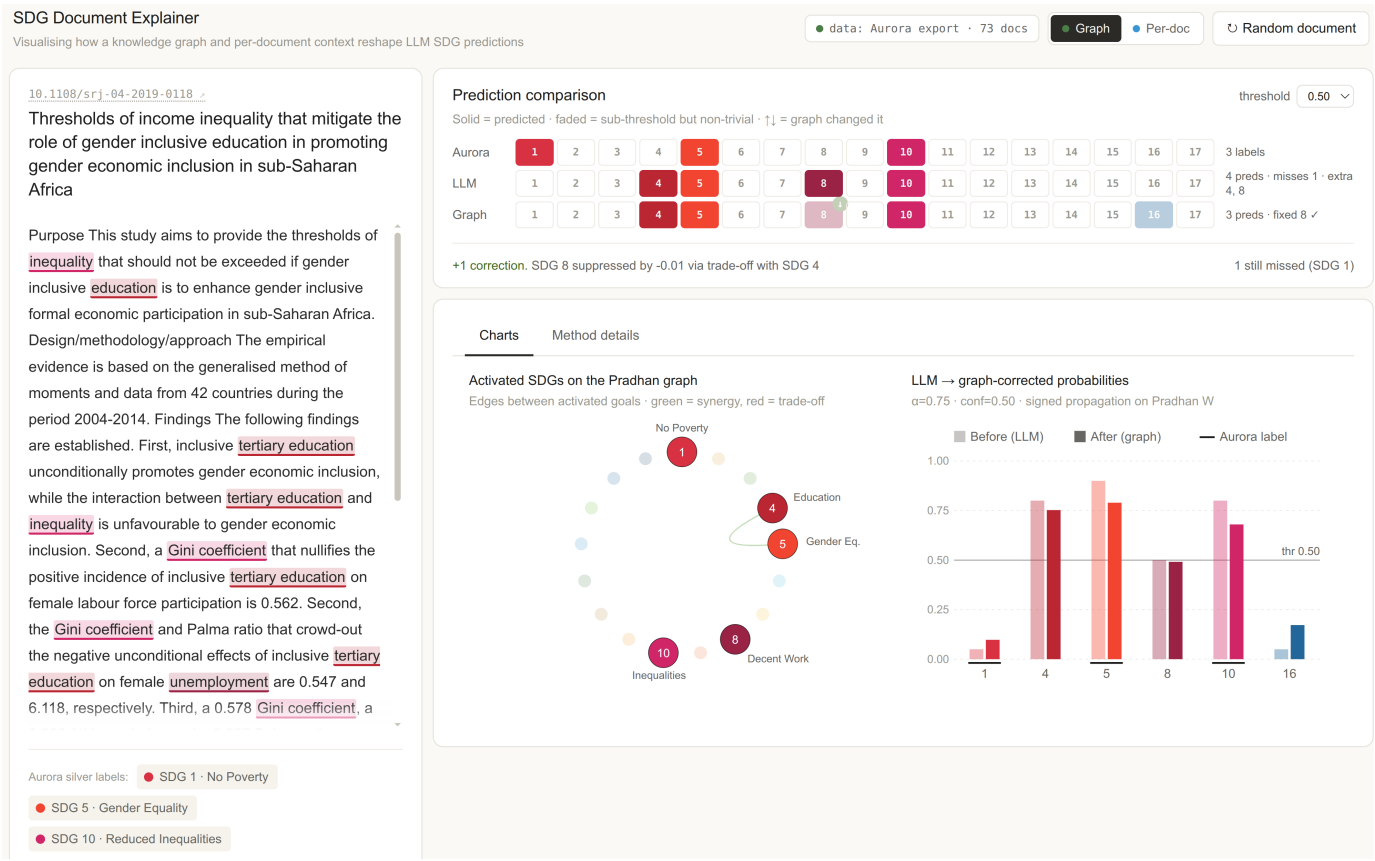


Fig. 2. The Explainer dashboard on the Aurora paper [19]. *Left:* document panel with SDG-coloured keyword spans and silver-label chips (SDG 1, 5, 10). *Top right:* the prediction-comparison grid; the LLM recovers SDG 5 and 10, adds false positives SDG 4 and 8, and misses SDG 1; the graph method removes only the borderline false positive SDG 8 and leaves the rest unchanged. *Bottom right:* the activated subgraph and the paired LLM-vs-method probability bars.

B. Lenses

A toggle in the header switches between the graph and per-document correction methods, and every dependent panel re-renders to match. The dataset is loaded by a JavaScript file (`aurora_demo.js`); if absent, the explainer falls back to a small set of embedded demonstration documents, and a data-source badge in the header displays which source is active and how many documents are available. The contribution that distinguishes our explainer from existing per-document audit tools is the **drill-in view**, activated when the user clicks any SDG anywhere in the dashboard. Clicking pins the SDG across all panels: matching keyword spans glow, the selected column in the comparison grid is outlined, the network node grows and reveals its full set of Pradhan edges (including edges to currently-inactivated SDGs, so the user sees the *structural* profile rather than just the *document-specific* one), and matching probability bars stay full-opacity while non-matching ones dim. The *Method details* tab then switches from its global view to a three-column SDG-specific drill-in titled *"Why is SDG j hard to separate?"*, exposing three independent evidence sources for class entanglement.

1) **Lexical lens.** The keyword spans in this abstract that match the pinned SDG, each annotated with the *other* SDGs that the same span tags; co-tagging spans are direct evidence of vocabulary overlap. A summary line names the three SDGs whose keyword dictionaries have the highest Jaccard overlap with the pinned SDG.
2) **Corpus-correlation lens.** The full row of W for the pinned SDG, sorted by absolute strength, rendered as signed bars centred on zero (green for synergy, red for trade-off). W is by default an empirical Pearson correlation matrix of the 17 binary label columns across the loaded corpus, and is therefore sample-sized on a small export; substituting a published interlinkage matrix such as Pradhan's [3] removes this dependence, with identical visualisation logic.
3) **Co-label lens.** Two stacks of bars: the first reports $P(\text{SDG } j \mid \text{pinned SDG})$ across all documents in the loaded dataset; the second counts, among the document's top-5 SBERT-nearest neighbours that carry the pinned SDG, how often each other SDG is also carried. The first is a corpus-wide property; the second is local to the current document.

The co-label lens is a static, corpus-wide label-frequency statistic over the loaded dataset's gold annotations; it does not condition on the document's text and is not used as a classifier, its role being purely to surface annotator behaviour for comparison with the other two lenses. The three lenses are deliberately heterogeneous: lexical overlap is a property of the

keyword dictionary and of the specific text, Pradhan synergy is a property of UN indicator time series at the country level with no exposure to text, and co-label rate is a property of human annotation behaviour. We read convergence across the three lenses as a signature consistent with political construction rather than as proof of it.

C. Example

We walk through the drill-in on a representative paper: Asongu and Odhiambo's study of income-inequality thresholds and gender-inclusive education in sub-Saharan Africa [19], with SDG 5 (Gender Equality) pinned. The Aurora silver labels are SDG 1 (No Poverty), SDG 5, and SDG 10 (Reduced Inequalities). The LLM correctly produces SDG 5 and SDG 10 with high confidence, but adds two false positives—SDG 4 (Quality Education, on a heavy "tertiary education" vocabulary) and SDG 8 (Decent Work, on "female labour force participation" / "unemployment")—and *misses* SDG 1 entirely. The graph-correction method removes only the borderline false positive SDG 8 (its probability falls from 0.50 to 0.49) and otherwise leaves the prediction intact—it still carries the SDG 4 false positive and still misses SDG 1—a useful negative example showing that signed propagation alone is insufficient when the document's vocabulary itself misleads the model.

Read across the three lenses, the picture is consistent. The **lexical lens** is locally sparse: the only span tagging SDG 5 in the abstract is the single word "women", the rest being inequality and education vocabulary ("Gini coefficient", "Palma ratio", "tertiary education", "unemployment"), so on text alone SDG 5 looks weakly supported. The **corpus-correlation lens** complicates that reading: SDG 5's row shows strong synergy with SDG 16 (Peace & Justice, $+0.40$) but essentially no linear correlation with SDG 10 (≈ 0), alongside trade-offs with the environmental cluster (SDGs 11–15), placing it inside a social-policy cluster and apart from environmental goals. The **co-label lens** recovers the SDG 10 link the correlation lens misses: among the loaded papers the conditional rates $P(j \mid 5)$ are SDG 16 57%, SDG 10 43%, SDG 1 14%, and among the SBERT neighbours sharing SDG 5 the labels 1, 4, and 16 each recur in half the cases. That a strong co-occurrence (SDG 10, 43%) coincides with a near-zero linear correlation is exactly the kind of cross-lens disagreement the tool is built to surface. These are small samples, computed over the SDG 5 papers within the loaded export rather than the full corpus, so we read the convergence as suggestive rather than conclusive.

Figure 3 contrasts the two correction methods on this same paper through the *Method details* view, and makes their different mechanisms concrete. Both eliminate the SDG 8 false positives, but by different routes and with different collateral effects. The graph method (top) nudges SDG 8 just below threshold ($0.50 \rightarrow 0.49$) and otherwise leaves the prediction untouched; its "strongest edges" panel shows that *none* of the activated goals are linked above $|W| \geq 0.10$, so on this document the correction reduces to an α-shrinkage with no structural signal to exploit—a per-document illustration of why no graph variant beats the baseline in §III. The per-document method (bottom) instead reweights each goal by SBERT-neighbour and entity evidence: it removes SDG 8, but the same reweighting also drives the *true* label SDG 10 below threshold, trading a false-positive fix for a false-negative regression. The two panels thus localise, for one document, the precision–recall tension that the aggregate macro-F_1 numbers only summarise: structural correction is too weak to act, while document-conditional correction helps on average yet can suppress correct labels on individual documents. The explainer's value here is diagnostic—it shows *which* evidence source moved *which* goal, rather than reporting a single corrected score.

The failure mode is now legible: the LLM did not extract a sufficient SDG 1 (poverty) signal because the abstract uses quantitative terms such as "Gini coefficient" and "Palma ratio" rather than the canonical poverty vocabulary. The wider point is not that SDG 1 was missed here, but that across the three lenses the boundaries between SDG 5, SDG 10, SDG 16, and SDG 1 are not drawn at natural joints of the data: the 2015 taxonomy assigned each a distinct goal number to satisfy political constituencies, and the empirical research record provides limited support for treating them as four cleanly separable things.

D. Discussion

The three lenses provide sufficient information to examine the blurred class boundaries between SDGs. It is equally important to understand the limitations of the current method.

a) What the explainer does not *do:* The drill-in does not assert that any specific SDG pair is misconceived. It surfaces quantitative agreement among three independent evidence sources; the interpretive step (e.g., political-construction artefact, indirect causal pathway, silver-label noise, benchmark or corpus-composition bias, or residual model limitation) is left to the user. The lenses are useful precisely because they refuse to commit to a single reading.

b) Implications for ML practice: If a substantive fraction of the residual macro-F_1 gap in SDG classification is class entanglement rather than calibration error [20], then several mainline strategies for closing it will continue to disappoint: per-class thresholds cannot separate classes the labels do not separate, and larger LLMs will not learn boundaries the training labels do not draw. Our corrected evaluation is consistent with this diagnosis: no graph-correction variant exceeds the no-correction baseline, which is what one expects when the activated classes are barely linked in the interaction matrix—signed propagation then has no structural signal to exploit, and the gain instead comes entirely from document-conditional neighbour evidence. The honest interventions are at the label-design level—hierarchical or soft labels, ordinal interlinkage codes such as the Nilsson 7-point scale [5], [21], or a target-level formulation that forces the political aggregation to be explicit. Our tool supports that work by giving domain experts a per-document interface to inspect entanglement candidates.

c) Generality: The explainer is parametrised on three inputs: a class set, a class-by-class interaction matrix, and a

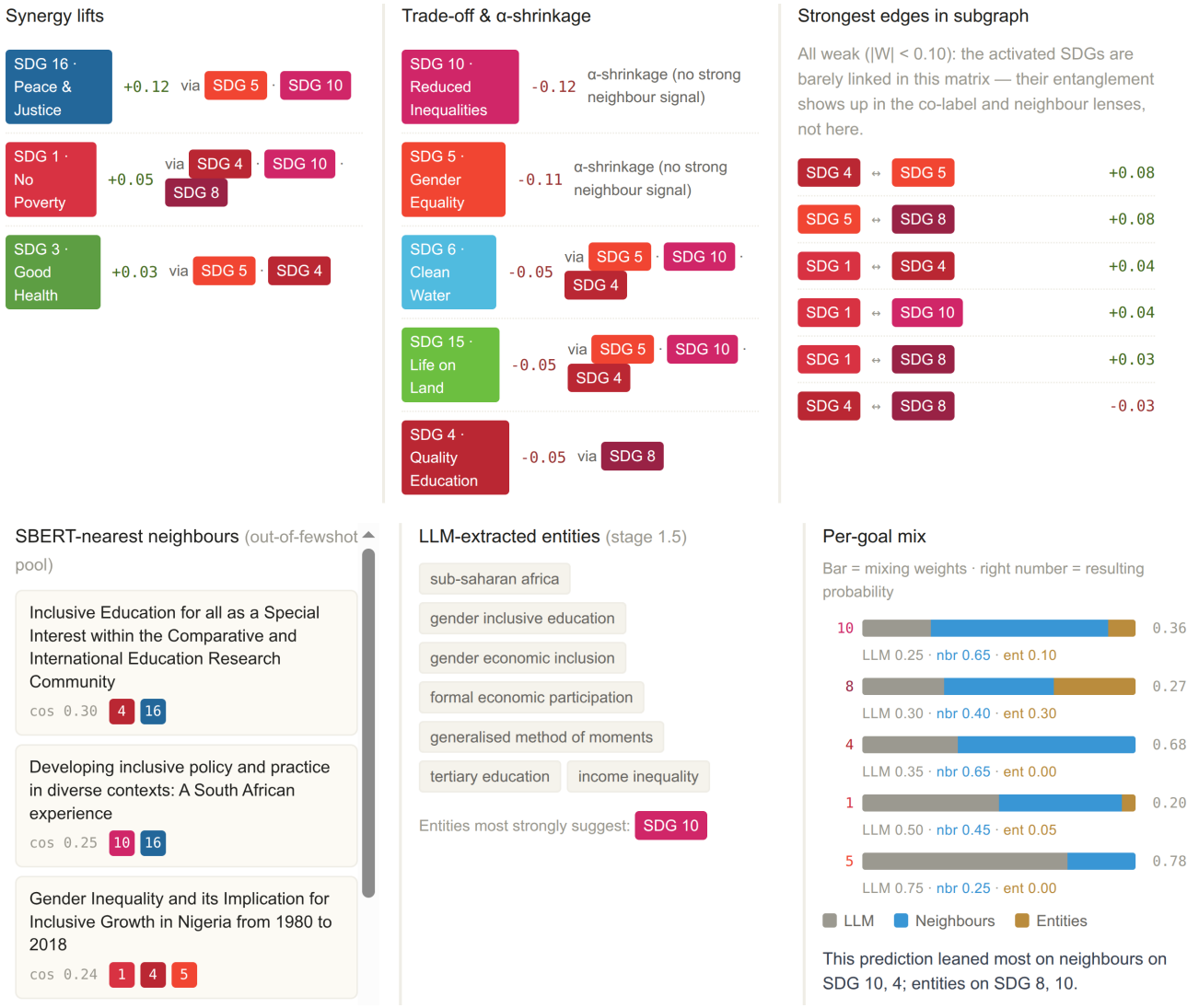


Fig. 3. The Methods panel showcasing two correction methods on the same paper (Asongu & Odhiambo [19], threshold 0.5). *Top:* graph correction (signed propagation on the corpus-correlation matrix) drops the borderline false positive SDG 8 and keeps SDG 10. *Bottom:* per-document correction (neighbour and entity evidence) also fixes SDG 8 but suppresses the true label SDG 10. Both remove the SDG 8 false positives, yet they target different failure modes.

per-class keyword dictionary. With no code changes, it applies to any multi-label classification problem with a negotiation-based taxonomy and a published interlinkage matrix (e.g., IPTC NewsCodes for news classification, or EU policy classification on EUR-Lex). The same per-document, label-structure-first perspective also transfers to other large multi-label text settings where labels routinely co-occur, such as clinical ICD coding [22]. In practice, the tool is a single self-contained HTML file with no build step, server, or runtime dependencies: a dataset is a generated JavaScript export dropped beside it, and an embedded fallback set of documents makes it runnable with zero configuration. We plan to release the explainer together with its export schema so that any compatible SDG (or other taxonomy) pipeline can drive it.

d) Coverage: Our keyword dictionary, derived from UN indicator metadata, is fixed across documents; a learned per-document key-phrase extractor would sharpen the lexical lens. Corpus-level co-label statistics are computed only over the loaded export, not the full Aurora corpus, and when a Pradhan-style interlinkage matrix is substituted, it is itself an approximation and is contested [17].

e) Evaluation under overlapping classes: The standard multi-label setup treats any prediction outside the gold set as a false positive by construction. For taxonomies designed with deliberate overlap—of which the SDGs are one instance—this can understate performance: an off-gold prediction may be a valid but unlabelled class rather than a model error. We do not resolve which holds in any given case; with Aurora's *silver*, query-assigned labels, an off-gold prediction is equally consistent with a missing-but-valid label, label noise, or a true error. The contribution here is to make those competing explanations inspectable per document.

V. Conclusion

We have presented an interactive explainer for multi-label SDG document classification that surfaces, on demand, three independent lenses on class entanglement (lexical, corpus-correlation, and co-label). On a representative paper from the Aurora corpus, all three lenses converge on the same conclusion: the SDG that the model missed showed little separation, by every measure available here, from the SDGs it correctly produced. The explainer reframes the residual error in SDG classification as a class-design question rather than a model question, and provides a per-document interface for the class redesign needed to address it.

Two extensions follow naturally from the current design. First, a controlled user study with sustainability domain experts would quantify whether the three-lens drill-in actually shifts expert beliefs about which classifier "errors" are diagnostic of model deficiency versus class-design artefacts. Second, we plan a target-level (169-class) version of the drill-in that exposes the within-goal aggregation explicitly: the choice to roll target 5.6 (reproductive rights) and target 5.5 (women in leadership) into a single SDG 5 prediction is itself a political aggregation worth surfacing.

Acknowledgments

This work has received funding from the Austrian Research Promotion Agency (FFG) through the GreenGLAM (GA No. 910266) and CLAIM (GA No. 923561) projects in the AI for Green and AI Ecosystems programmes, respectively, and from the European Union's Horizon Europe programme through the BENEFITS project (GA No. 101179032).

References

[1] United Nations General Assembly, "Transforming our world: The 2030 agenda for sustainable development," Resolution A/RES/70/1 adopted 25 September 2015, 2015. [Online]. Available: https://sdgs.un.org/2030agenda

[2] M. Vanderfeesten, R. Otten, and E. Spielberg, "Search queries for "Mapping research output to the Sustainable Development Goals (SDGs)" (version 5.0)," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3817445

[3] P. Pradhan, L. Costa, D. Rybski, W. Lucht, and J. P. Kropp, "A systematic study of Sustainable Development Goal (SDG) interactions," *Earth's Future*, vol. 5, no. 11, pp. 1169–1179, 2017. [Online]. Available: https://doi.org/10.1002/2017EF000632

[4] D. Le Blanc, "Towards integration at last? the sustainable development goals as a network of targets," *Sustainable Development*, vol. 23, no. 3, pp. 176–187, 2015. [Online]. Available: https://doi.org/10.1002/sd.1582

[5] M. Nilsson, E. Chisholm, D. Griggs, P. Howden-Chapman, D. McCollum, P. Messerli, B. Neumann, A. Stevance, M. Visbeck, and M. Stafford-Smith, "Mapping interactions between the sustainable development goals: lessons learned & ways forward," *Sustainability Science*, vol. 13, pp. 1489–1503, 2018. [Online]. Available: https://doi.org/10.1007/s11625-018-0604-z

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[8] B. Alsallakh, A. Hanbury, H. Hauser, S. Miksch, and A. Rauber, "Visual methods for analyzing probabilistic classification data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1703–1712, 2014. [Online]. Available: https://doi.org/10.1109/TVCG.2014.2346660

[9] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and J. Suh, "ModelTracker: Redesigning performance analysis tools for machine learning," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, 2015, pp. 337–346. [Online]. Available: https://doi.org/10.1145/2702123.2702509

[10] Institute for Global Environmental Strategies, "SDG interlinkages analysis and visualisation tool (V4.0)," Online tool, 2023. [Online]. Available: https://sdginterlinkages.iges.jp/

[11] A. Scharl, D. Herring, W. Rafelsberger, A. Hubmann-Haidvogel, R. Kamolov, D. Fischl, M. Föls, and A. Weichselbraun, "Semantic systems and visual tools to support environmental communication," *IEEE Systems Journal*, vol. 11, no. 2, pp. 762–771, 2017. [Online]. Available: https://doi.org/10.1109/JSYST.2015.2466439

[12] C. S. Armitage, M. Lorenz, and S. Mikki, "Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?" *Quantitative Science Studies*, vol. 1, no. 3, pp. 1092–1108, 2020. [Online]. Available: https://doi.org/10.1162/qss_a_00071

[13] P. J. Purnell, "A comparison of different methods of identifying publications related to the United Nations Sustainable Development Goals: Case study of SDG 13—climate action," *Quantitative Science Studies*, vol. 3, no. 4, pp. 976–1002, 2022. [Online]. Available: https://doi.org/10.1162/qss_a_00215

[14] R. Mutz, L. Bornmann, and R. Haunschild, "How to use assignments of United Nations Sustainable Development Goals (SDGs) to scientific papers in research evaluation? the proposal of a gold standard combining assignments from different data providers," *Scientometrics*, vol. 130, pp. 1519–1546, 2025. [Online]. Available: https://doi.org/10.1007/s11192-025-05254-w

[15] T. Matsui, K. Suzuki, K. Ando, Y. Kitai, C. Haga, N. Masuhara, and S. Kawakubo, "A natural language processing model for supporting sustainable development goals: Translating semantics, visualizing nexus, and connecting stakeholders," *Sustainability Science*, vol. 17, no. 3, pp. 969–985, 2022. [Online]. Available: https://doi.org/10.1007/s11625-022-01093-3

[16] I. Rafols, E. Noyons, H. Confraria, and T. Ciarli, "Visualising plural mappings of science for Sustainable Development Goals (SDGs)," SocArXiv preprint, 2021. [Online]. Available: https://doi.org/10.31219/osf.io/yfqbd

[17] A. Warchold, P. Pradhan, and J. P. Kropp, "Variations in sustainable development goal interactions: Population, regional, and income disaggregation," *Sustainable Development*, vol. 29, no. 2, pp. 285–299, 2021. [Online]. Available: https://doi.org/10.1002/sd.2145

[18] J. E. Guisiano, R. Chiky, and J. De Mello, "SDG-Meter: A deep learning based tool for automatic text classification of the Sustainable Development Goals," in *Intelligent Information and Database Systems (ACIIDS 2022)*, ser. Lecture Notes in Computer Science, vol. 13757. Springer, 2022, pp. 259–271. [Online]. Available: https://doi.org/10.1007/978-3-031-21743-2_21

[19] S. A. Asongu and N. M. Odhiambo, "Thresholds of income inequality that mitigate the role of gender inclusive education in promoting gender economic inclusion in sub-Saharan Africa," *Social Responsibility Journal*, vol. 17, no. 1, pp. 106–126, 2021. [Online]. Available: https://doi.org/10.1108/SRJ-04-2019-0118

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[21] International Council for Science, "A guide to SDG interactions: From science to implementation," ICSU, Paris, Tech. Rep., 2017. [Online]. Available: https://council.science/publications/a-guide-to-sdg-interactions-from-science-to-implementation/

[22] F. Teng, Y. Liu, T. Li, Y. Zhang, S. Li, and Y. Zhao, "A review on deep neural networks for ICD coding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 4357–4375, 2023. [Online]. Available: https://doi.org/10.1109/TKDE.2022.3148267