

From Signals to Barriers: An LLM-Assisted Bow-Tie Framework for NGO Civic-Space Risk Monitoring

Noemí Orué-Guerrero¹ Bernat Coma-Puig²

¹Department of Political Science, Constitutional Law and Philosophy of Law,
University of Barcelona, Barcelona, Spain

²Independent Researcher, Barcelona, Spain

coruegue8@alumnes.ub.edu

bernat.coma.puig@gmail.com

Abstract

Non-governmental organisations (NGOs) increasingly operate in environments where regulatory, financial, administrative, and reputational pressures shift fast and unpredictably. The sector's standard risk tools (registers, likelihood-impact matrices, key risk indicators) might help to analyse existing risks, but they do not expose the causal chain from an early signal to its operational consequence, and they do not stay updated without maintenance that under-resourced NGO staff simply cannot provide. We tackle both problems. First, we adapt the bow-tie methodology from safety-critical industries to the NGO context, anchored on a single top event: the loss of operational viability in a country, with five threat categories and explicit barriers. Second, we instantiate an LLM-assisted maintenance system that classifies incoming signals, updates a versioned, evidence-grounded knowledge base, and flags barriers for human verification, while leaving consequential decisions to people. We illustrate the framework with a worked example from public reports (CIVICUS, ICNL), showing structured, transparent decision support as a concrete form of positive impact for NLP applied to civic resilience.

1 Introduction

Non-governmental organisations (NGOs) now operate in far more demanding regulatory environments than they did two decades ago. More than sixty countries adopted measures restricting foreign funding to NGOs between 1994 and 2015 (Bromley et al., 2020), and the trend has not reversed: 92% of the world's population now lives in countries where civic space, that is, the degree of protection afforded to the freedoms of association, peaceful assembly and expression, is rated as obstructed, repressed or closed (CIVICUS, 2025). Restriction rarely takes the form of outright prohibition. It arrives as an accumulation of legal,

administrative, financial and reputational measures applied through ordinary channels of government, each plausibly technical when taken on its own: an additional registration requirement, a more frequent audit, a change to the bank authorised to receive funds from abroad (Carothers and Brechenmacher, 2014; DeMattee, 2019). Their restrictive function only becomes legible when the measures are read together and over time.

Signals of change, such as draft bills, regulatory circular, hostile official statements, and tighter banking practices, are scattered across sources and languages, in volumes no compliance team can track. The operational consequence, moreover, is delayed: a measure that creates control capacity does not restrict from the moment it is adopted. In Uzbekistan, the obligation to channel foreign assistance through the state bank closed no organisation by itself, but later allowed the government to block more than 80% of transfers to local NGOs (Carothers and Brechenmacher, 2014). That lag is the anticipation margin, and it passes unnoticed because it does not yet look like a restriction. The difficulty also varies by organisation: an international NGO must watch signals in remote jurisdictions and languages, while a locally rooted organisation is closer to them but rarely has the legal or compliance capacity to follow them systematically.

What is missing, then, is a representation that connects threats, the loss of control they can produce, the barriers standing in the way, and the consequences that follow, plus a mechanism that keeps that representation current without adding yet another monitoring burden onto already overstretched staff.

The bow-tie method covers the first half of this gap. Developed in safety-critical industries, it organises a risk around a central top event (the moment control is lost), with threats and preventive barriers on one side and consequences and recovery barriers on the other. It is explicitly causal and

explicitly operational: each threat line names the defences that are supposed to hold, so a change on that line points straight to a concrete barrier to check. The method has started to appear in non-profit risk practice as a response-planning and KRI-design aid (GRF CPAs & Advisors, 2019; Wares, 2021), but in practice, applications remain occasional workshop exercises that produce static diagrams rather than live representations of an organisation’s threat situation.

We adapt the bow-tie to the civic-space problem by anchoring it on a single top event: the *loss of operational viability in a country*, the point at which the accumulated restrictiveness of the environment exceeds what delivering the core mandate requires. This reflects a pattern documented in the civic-space literature, in which restrictions often accumulate through legal and administrative measures rather than through a single act of prohibition (DeMattee, 2019); deregistration, asset freezes, and expulsion are consequences that follow once viability is lost, not separate top events in their own right. We organise the threat side into five categories: four from DeMattee (2019)’s typology of regulatory provisions, plus reputational and political delegitimation drawn from the closing-civic-space literature (Carothers and Brechenmacher, 2014).

The second half of the gap is keeping the diagram current. LLMs have shown promise for horizon scanning and policy-document classification (Zhang et al., 2025; Hassani et al., 2025; Gunes and Florczak, 2023), and NLP has supported human-rights monitoring (Alhelbawy et al., 2020). Related NLP4PI work has applied LLMs to policy monitoring (Lu, 2025), emphasised deployment and maintenance in humanitarian settings (Abilov et al., 2025), and highlighted the need for stakeholder participation (Wilson et al., 2025). We differ by treating monitoring output as a maintained, versioned causal risk representation rather than a stream of isolated classifications. The LLM acts as a maintenance layer: it classifies signals, updates an organisation-specific knowledge base and bow-tie rendering, and surfaces barriers for human verification, while consequential decisions remain with people.

We illustrate the framework with a worked example drawn from public reports (CIVICUS, ICNL), in which a language model performs the retrieval and filing roles for that case, and render the resulting bow-tie views for inspection. We propose this combination, an established causal risk repre-

sentation designed for continuous LLM-assisted maintenance under an auditable editorial policy, as a concrete direction for NLP applied to civic resilience.

2 The NGO Bow-Tie Framework

2.1 Hazard and Top Event

The bow-tie method comes from process-safety engineering. The earliest documented form of bow-tie analysis is commonly traced to hazard-analysis course material prepared by Imperial Chemical Industries for a 1979 lecture at the University of Queensland (Alizadeh and Moshashaei, 2015). Shell subsequently incorporated the method into its risk-management practices in the early 1990s, in the context of organisational changes following the Piper Alpha disaster (Wolters Kluwer). A bow-tie represents a *hazard* together with a *top event*: the point at which control over that hazard is lost, with *threats* and *preventive barriers* on the left and *consequences* and *recovery barriers* on the right.

We define the hazard as *operating as an NGO within a sovereign state’s regulatory regime under conditions of contested legitimacy*, and the top event as the *loss of operational viability in a country*: the point at which the accumulated restrictiveness and complexity of the regulatory environment exceed what the organisation needs to deliver its core mandate. We treat this as a single top event, rather than setting up separate events for expulsion, deregistration, or asset freezes, for three reasons: i) it matches the dominant empirical pattern: states constrain civil society less through one dramatic act than through the incremental build-up of restrictive provisions applied via ordinary administrative channels (DeMattee, 2019; Carothers and Brechenmacher, 2014), since organisations are usually strangled by friction rather than formally expelled; ii) expulsion, deregistration, and asset seizure are not independent events but *consequences* that follow once viability is already lost; collapsing them into one top event keeps the threat side and the consequence side distinct; and iii) it gives the downstream system (Section 3) a single, well-defined target that every signal routes into.

2.2 Threat Taxonomy

We organise threats into five categories. Four come from DeMattee (2019)’s typology of the regulatory provisions through which states govern civil society: *governance*, *formation*, *operations*, and

resources. The fifth, *reputational and political delegitimisation*, captures the non-legal pressures documented in the closing-civic-space literature (Carothers and Brechenmacher, 2014; Dupuy et al., 2016) that fall outside the four regulatory categories but often precede or accompany them. Table 1 lists all twenty subcategories.

The taxonomy has a structural feature the system relies on. In DeMattee’s framework, each restrictive provision has a permissive counterpart that protects the organisation: a right of appeal against arbitrary deregistration, say, or a statutory deadline on a registration decision. In bow-tie terms, the restrictive provision is a *threat* and its permissive counterpart a *preventive barrier*. So a classifier that detects whether a document instantiates a restrictive or a permissive provision is, in effect, detecting whether a threat has fired or a barrier is holding.

2.3 Barrier Chains

On the threat side of the bow-tie (left of the top event in Figure 1), each threat is connected to the top event through *preventive barriers*, populated from the permissive provisions introduced above (Section 2.2): appeal rights and judicial dispute resolution (governance); statutory deadlines and bounded grounds for refusal (formation); general operating licences and recognised self-regulation (operations); diversified, partly domestic funding with transparent reporting (resources). Against reputational threats, the barriers are non-legal: proactive transparency, and the public trust that permissive provisions are meant to build. On the consequence side of the bow-tie (right of the top event), the top event is connected to its *consequences* (programme suspension, loss of beneficiary access, endangerment of local staff, loss of mandate or funding) through *recovery barriers*: contingency and relocation planning, pre-arranged handover to local partners, staff evacuation procedures, and financial reserves paired with proactive donor communication. Figure 1 renders the complete diagram. Keeping the barrier set explicit and finite is what lets a signal, once classified into a given subcategory, point to exactly which barriers need verification, rather than raising an undifferentiated alert.

2.4 Consequences and Recovery Barriers

The right side of the bow-tie is comparatively generic. Where the threat taxonomy (Section 2.2) is anchored to the civic-space literature on NGO regulatory provisions, the four consequences and

their recovery barriers (contingency and relocation planning, local-partner handover, staff evacuation, and reserves paired with donor communication) draw on standard NGO operational-continuity and security practice. This is by design: what happens once viability is lost, i.e., programmes stopping, beneficiaries losing access, staff at risk, funding drying up, is largely the same whichever threat caused it, so a single generic recovery structure suffices. The paper’s contribution is concentrated on the threat side, i.e., the threat taxonomy and the mechanism that keeps it current; we include the right side (Figure 1) so the bow-tie reads as a closed diagram.

3 The System, Illustrated

The framework in Section 2 describes an organisation’s threat situation at a given moment. In practice, though, that situation changes over time: signals build up, earlier measures take on new meaning, and how relevant a document is often depends on what was already known when it appeared. This section describes a continuous system that maintains a living, evidence-grounded assessment: it collects signals as they appear, reads each new entry against the existing record, updates a versioned knowledge base, and renders the result in a form a person can read, verify, and correct.

3.1 Architecture

We describe the system as an architecture (Figure 2) rather than an implementation: components are defined by role, leaving technology choices open. The system’s product is not a stream of alerts or on-demand regenerated assessments, but a maintained, versioned knowledge base co-authored by machine and human, from which alerts and diagrams are derived.

The collector. Entries enter the system through a component that deliberately exercises no judgement. The collector works from a *source registry*: a configured, versioned list of what to watch, i.e., legislative portals and gazettes, regulator and central-bank pages, official statements and court filings, the reporting of civic-space monitors such as the CIVICUS Monitor and the ICNL trackers, and selected news sources. On each run it fetches what is new or changed, normalises it to text, stamps it with source, date, jurisdiction, and document type, and appends it to the store. Whether a document *matters* is not its concern, that is the maintainer’s

Subcategory	Example signal
Governance <i>Provisions that create and empower the regulator and constrain constitutionally protected freedoms.</i>	
Unchecked regulator powers	New agency empowered to seize NGO property or documents without cause
Loss of appeal rights	Law removes the right to appeal a deregistration order
Restricted freedoms	Constitutionally protected freedom of association curtailed by statute
Unilateral rule-making	Minister empowered to issue new NGO rules without legislative process
Formation <i>Provisions governing whether and how organisations may legally come into existence.</i>	
Forced (re)registration	Mandatory re-registration triggered by renaming or constitutional change
Arbitrary denial	Registration refused for reasons not stated in law
Capital / membership bars	Punitive minimum capital or founder-count requirements
Mandatory state partner	Registration conditioned on a government sponsor or MoU
Operations <i>Provisions controlling what activities an organisation may carry out and how.</i>	
Per-project permits	Separate government permit required for each individual project
Foreign-funding activity bans	Activities (e.g. advocacy) prohibited if foreign funds exceed a threshold
Forced coordination	Agency interference in, or compelled coordination of, operations
Asset surrender	NGO required to surrender operational assets to the state on dissolution
Resources <i>Provisions regulating how an organisation may receive, hold, and report funds.</i>	
Foreign-funding approval	One-time or per-transaction state approval required to receive foreign funds
Foreign-funding caps / bans	Ceiling on, or prohibition of, foreign funding from certain sources
Mandatory state banking	Funds must be held in designated government banks
Punitive taxation	Foreign funds taxed where comparable domestic funds are not
Reputational & political <i>Non-legal pressures that erode public trust and frame the organisation as illegitimate.</i>	
“Foreign agent” labelling	Official designation stigmatising foreign-funded organisations
Hostile official rhetoric	Public statements by officials framing NGOs as threats to sovereignty
Coordinated media campaigns	Sustained negative media targeting an organisation or the sector
Criminalising association	Public accusations linking NGOs to espionage or subversion

Table 1: Taxonomy of threats to NGO operational viability. The first four regulatory categories and their subcategories are adapted from [DeMattee \(2019\)](#); the reputational and political category is drawn from the closing-civic-space literature ([Carothers and Brechenmacher, 2014](#)). Several resources and formation subcategories concerning foreign funding also draw on [Dupuy et al. \(2016\)](#) and [Bromley et al. \(2020\)](#).

job. The registry is not limited to public sources: it includes a manual inlet through which authorised staff deposit the organisation’s own private correspondence (a warning letter from a ministry, a bank’s notice of an account restriction, a call summary from counsel), a source whose feed is a person, and often where the earliest and most consequential signal actually originates. Private entries are tagged with sensitivity labels and access permissions at ingestion time, so they can support internal assessments without being exposed in exported reports. Keeping the registry explicit and versioned means the question “*why did the system not see this?*” always has an auditable answer: either the source was never registered, or the entry is in the store and the maintainer’s filing of it can be inspected.

The evidence store. What the collector deposits is never touched again. The store is the system’s

memory and its audit trail: an append-only record of every entry, each carrying its source, date, jurisdiction, and sensitivity metadata. Nothing is discarded, because a signal that looks minor today is what gives next month’s signal its meaning, and every claim the system later makes has to resolve back to dated entries here. A staff entry sits on the same level of importance as a public newspaper entry (dated, sourced, citable) even though it will never appear in a public monitor and may carry different access rights. The store is never edited, only cited.

The knowledge base. The second component is the product: a bounded set of *articles*, one per threat line (per jurisdiction, where an organisation operates in several), plus a country overview. Each article carries three things: structured state, including the line’s status, one of *unassessed*, *low risk*, *high risk*, or *materialised*, together with its open

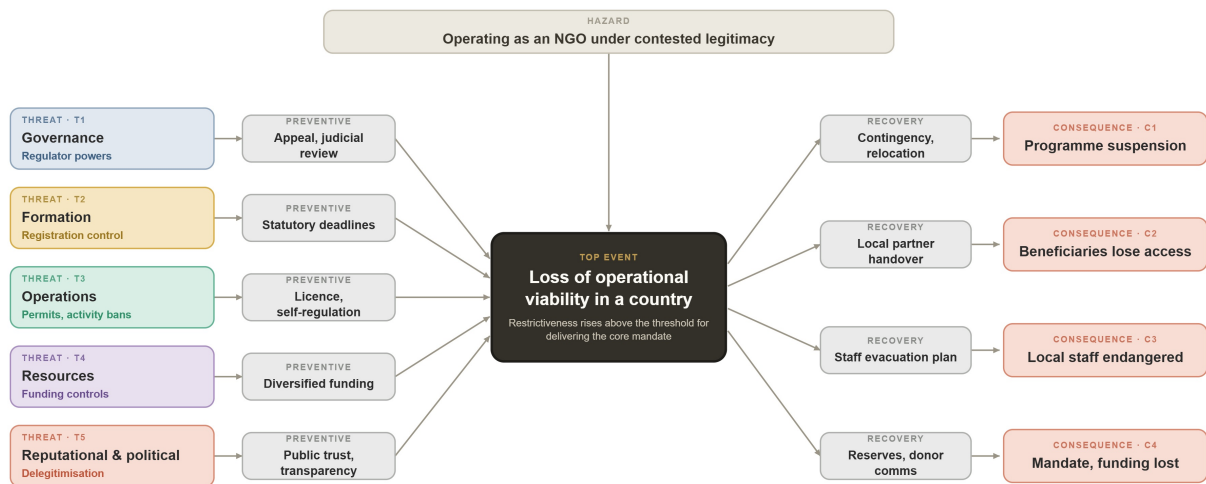


Figure 1: The NGO bow-tie: five threat categories and their preventive barriers on the left, converging on the top event (loss of operational viability in a country), with consequences and their recovery barriers on the right.



Figure 2: System architecture. The collector feeds an append-only evidence store; the maintainer, an LLM agent operating under a versioned editorial policy, reads the store and writes the knowledge base; humans can also edit the knowledge base directly. All outward-facing renderings (the periodic e-mail, the bow-tie diagram) are derived from knowledge-base state.

items and the date last assessed; a narrative account of how the line reached that state; and citations linking every substantive claim to dated entries in the evidence store. *Unassessed* denotes insufficient evidence coverage, while *low risk* requires affirmative permissive evidence. Versioning is part of the data model, not a convenience of tooling: every change to an article is recorded with its date, its author (machine or human), and its rationale. This single design decision answers the three questions an organisation actually asks. The current state of an article answers *where do we stand*; its revision history answers *how did we get here*; the difference between two revisions answers *what changed since I last looked*.

The maintainer. The articles described above are produced and kept current by the third component: an agent whose job is to read the evidence store

and turn it into knowledge. Like any agent, it needs instructions to operate reliably. These take the form of a versioned *editorial policy*, independent of any particular model, which fixes the article granularity, the allowed statuses, how much evidence is needed to move between them, how uncertainty is treated, the requirement that every claim be cited, and the rule that human edits always take precedence over the agent's own. Under this policy, a large language model performs two operations. It ingests: each arriving entry is assessed for relevance and may receive zero or more category-specific assignments, each classified as *restrictive*, *permissive*, or *contextual*. Irrelevant records receive no assignment. Because edits are made against the current article rather than regenerated from scratch, assessment is inherently cumulative: the account can say that funding restrictions have tightened across several

instruments over months, rather than flagging today’s news source entry as an isolated alert. It also *audits*: a periodic pass over the knowledge base itself, checking for internal contradictions, claims a newer document supersedes, unresolved open items, and statuses the cited evidence no longer supports. The audit exists because a knowledge base that asserts statuses decays dangerously: an old, unrevised *low risk* masks changed conditions, and an abandoned *unassessed* line creates false reassurance from silence. The contract is the same across both operations: every edit is bounded, cited, reversible, and justified in its revision record.

Automated maintenance and co-authorship.

The architecture automates scheduled retrieval, ingestion, auditing, and rendering, but not organisational decision-making. Decisions such as closing a programme, relocating staff, or activating a recovery barrier remain human responsibilities. The PoC (see Appendix A) demonstrates that the maintenance stages are executable, not that they operate reliably without supervision over time. Citations make unsupported claims inspectable, versioning makes revisions reversible, and periodic reports expose every change to human review. Human edits and machine edits are recorded separately, preserving disagreement rather than overwriting it.

3.2 Operation

Section 3.1 describes what the system is built from. This subsection describes how it behaves over time: how its state reaches a person, on what schedule it updates, and what it takes to build one today.

Delivery. Users receive the updated bow-tie through an appropriate delivery channel, such as e-mail, a dashboard, or an internal reporting system, together with a citation-backed summary of changes since the previous run. For each status transition, the system explains what changed and why, so the movement can be understood without first opening the underlying article. The bow-tie is the canonical view of the assessment: its structure remains fixed, while each threat line displays its current status (*unassessed*, *low risk*, *high risk*, or *materialised*) and links to the narrative and dated evidence. Different users can therefore access the same assessment at different levels of detail. The knowledge base retains the full revision history, allowing past trajectories to be reconstructed, as illustrated in Section 4.

The loop. The components interact in a cycle, *retrieve, ingest, audit, render*, that runs on a schedule matched to the phenomenon: civic-space restriction moves over weeks, so a periodic run (weekly, say, with an out-of-cycle pass when a watched source updates) keeps the account current without manufacturing noise from sources that rarely change. The intended workflow reduces the need for continuous manual monitoring, which is important if the system is to assist rather than add burden to an under-resourced organisation.

Instantiation. Nothing above names a specific technology, deliberately so, but the architecture is practical today. The framework is model-agnostic with respect to the filing and maintenance components: these can be implemented with hosted proprietary LLMs or locally deployed open-weight models, provided that they can follow the editorial policy and produce the required structured outputs. Evidence collection is a separate component and does not require the LLM itself to provide native web-search capabilities; it may instead rely on provider-specific web-search tools, independent search APIs, self-hosted metasearch infrastructure such as SearXNG, or domain-specific crawlers for legal and regulatory sources such as LegalCrawler (SearXNG Team, 2026; Chalkidis, 2020–2022). This separation allows organisations to choose models and deployment strategies according to their technical capacity, privacy requirements, and infrastructure constraints. Hosted models can reduce local infrastructure requirements, while locally deployed models can provide greater control over data and deployment.

One natural instantiation is the recently popularised *llm-wiki* pattern (Karpathy, 2026): articles as plain-text files under version control, the editorial policy as the instruction file governing an LLM agent, edits committed directly to the revision history, the update generated from that history, and the bow-tie re-rendered on each run. The same architecture could also be implemented through a database-backed wiki with an editing interface for organisations whose staff do not work directly with repositories.

It differs from adjacent NLP systems in the way knowledge is treated over time. Existing work has generated grounded articles from source collections (Shao et al., 2024) and persisted structured summaries for subsequent querying (Edge et al., 2024), but does not describe the continuous, versioned,

audited, and human-co-authored maintenance of those outputs. In our setting, that maintenance process is not an auxiliary feature; it is the central function of the system.

4 Worked Example: Nepal, 2024–2026

We illustrate the architecture with a retrospective proof of concept on Nepal’s civic-space trajectory (2024–2026) from a foreign-funded INGO’s perspective. Prompts, evidence records, classifications, and outputs are provided in supplementary material.

Scope of the demonstration. The example is not tied to a real NGO, programme, or operational mandate and therefore lacks organisation-specific thresholds agreed with staff. It illustrates evidence filing, article revision, status transitions, and audit trails rather than providing an authoritative assessment of Nepal.

Experimental setup. The PoC comprised one retrospective collection job covering three assessment periods, batched evidence filing, and sequential article maintenance. Web access was enabled only during collection, which retrieved candidate evidence from ICNL, CIVICUS, and the open web without assigning categories, polarity, or status. The filing stage then classified each record from its title, description, and supporting snippet, flagging cases with incomplete grounding, poor taxonomy fit, ambiguity, or confidence below the untuned threshold of 0.70. The maintainer processed evidence chronologically, preserved supported state, and cited each substantive claim. Its four statuses were *unassessed*, *low risk*, *high risk*, and *materialised*. Low risk required affirmative permissive evidence, whereas insufficient evidence remained *unassessed*. The collector output was frozen for the final filing and maintenance run. Implementation details are provided in Appendix A. In our proof of concept, the complete pipeline incurred an API cost close to US\$1 per execution, despite no specific optimisation for cost. While this figure is only indicative and will vary with evidence volume, model choice, and retrieval strategy, it suggests that model and API costs are unlikely to constitute a major barrier to deployment for NGOs.

Baseline (31 December 2024). The retrospective backfill returned six records,¹ covering Nepal’s

CIVICUS rating, the TikTok ban and its reversal, statutory formation requirements, and foreign-funding approval rules. The reversal supported *governance: low risk*, while statutory requirements supported *formation: high risk* and *resources: high risk*. *Operations* and *reputational and political* remained *unassessed* because no sufficiently specific evidence supported a judgement.

Peak (late 2025). The collector returned 25 new records concerning draft NGO and social media legislation, the Media Council Bill, the temporary block of 26 social-media platforms and its reversal, the subsequent political crisis, and the loss of US-AID funding. The accumulated restrictive evidence raised *governance* from *low risk* to *high risk*, while *formation* remained *high risk*. *Resources* moved to *materialised* on the basis of the cumulative funding record, although donor-side evidence that did not fit the taxonomy exactly required human review. *Operations* and *reputational and political* remained *unassessed* because the former received no supported evidence and the latter only contextual material.

Divergence (June 2026). Here, *divergence* refers to the separation between political signals of improvement and the persistence of structural restrictions. The final window added nine records concerning the election and change of government, accountability measures following the 2025 crack-down, continued warnings about restrictive legislation, and a statement by the new Home Minister concerning media outlets. Although the political transition and some reversals provided permissive evidence, unresolved restrictive measures kept *governance* and *formation* at *high risk*, while *resources* remained *materialised*. *Operations* remained *unassessed*, and *reputational and political* also remained *unassessed* because the ministerial statement was insufficiently grounded to support a status change. No status changed in the final revision. This shows that permissive evidence does not automatically override unresolved restrictive evidence, and that weakly grounded signals do not move a threat line out of *unassessed*.

What this demonstrates, and what it does not. The pilot demonstrates the technical feasibility of automating the collection, classification, mainte-

¹Retrieval was performed in July 2026 using publication-date restrictions. This approximates, but does not reproduce,

a contemporaneous run because older material and search rankings may have changed. Complete artefacts are included in the supplementary material.

nance, and visualisation of evidence within a bow-tie risk framework. A real deployment would require the source registry, search strategy, taxonomy, and decision thresholds to be adapted to the organisation's mandate, countries of operation, funding model, and project portfolio. The present results therefore do not constitute an operational assessment of Nepal, but show that an LLM-assisted system can maintain a traceable and evolving risk representation from heterogeneous evidence.

5 Conclusion

We adapted bow-tie risk analysis to civic-space restriction, anchoring it on the loss of NGO operational viability in a country. Our contribution is threefold: an NGO-specific bow-tie representation, continuous rather than episodic monitoring, and, to our knowledge, the first LLM-assisted system for maintaining an NGO civic-space risk model, while consequential decisions remain human. The system files evidence, updates a versioned knowledge base, and surfaces changes.

A pilot on Nepal's 2024–2026 trajectory ran the architecture end to end on public evidence and exposed its gaps as clearly as its behaviour: a taxonomy that under-specifies digital-infrastructure threats, untested collector recall over multi-year backfill, and filing that still routes three quarters of records to human review.

We define positive impact more narrowly and verifiably than automation: widening the window between an early signal and its operational consequence, with every claim traceable to a dated source and every knowledge-base revision recorded and reversible. The next step is to measure that impact through filing accuracy against expert annotation, calibration of status transitions, and, ultimately, whether organisations act earlier than they otherwise would.

Limitations

This work presents an architecture and a retrospective proof of concept, not a deployed or quantitatively evaluated monitoring system. The Nepal example covers one country and an unusually well-documented period, and it was reconstructed after the relevant events had occurred. The authors therefore had knowledge of the eventual trajectory, which may make earlier signals appear more informative than they would have been in a prospective setting. The example demonstrates that the

proposed stages can be executed, but not that the system would identify the same evidence, classifications, or status transitions during live operation.

Ethical and Societal Implications

The principal societal benefit of this work is to help civil society organisations identify restrictive developments earlier and respond with better evidence. Relevant signals are often dispersed across legal documents, regulatory notices, monitoring reports, official statements, and different languages. By collecting and organising this material into a maintained risk representation, the system can reduce the monitoring burden on organisations with limited legal, compliance, or analytical capacity. Earlier and more structured awareness may give staff additional time to verify preventive barriers, consult local experts, communicate with donors, protect staff, or prepare operational alternatives before restrictions become critical.

The framework may also improve the transparency and continuity of organisational risk management. Each substantive claim is linked to dated evidence, changes are preserved through versioning, and the reasoning behind a status transition can be inspected. This can support institutional memory, reduce dependence on individual staff members, and make risk discussions more consistent across teams and reporting periods. The system is intended to support human judgement rather than replace it. Organisational staff remain responsible for interpreting the evidence and deciding whether any response is justified.

Several challenges must nevertheless be considered before deployment. Incorrect classifications may create unnecessary alarm, while missed evidence or outdated assessments may create false confidence. The knowledge base may also contain sensitive information, including private correspondence, organisational vulnerabilities, and contingency plans, which could place staff or partners at risk if accessed by hostile actors. In addition, public and official sources may reflect political or editorial perspectives, and the same monitoring framework could be adapted for surveillance of civil society. A real deployment would therefore require human review, data minimisation, appropriate access controls, secure hosting, sensitivity labels, and clear rules governing which information may be retained or shared.

References

- Anton Abilov, Ke Zhang, Hemank Lamba, Elizabeth M. Olson, Joel Tetreault, and Alex Jaimes. 2025. [Operationalizing AI for good: Spotlight on deployment and integration of AI models in humanitarian work](#). In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 189–195, Vienna, Austria. Association for Computational Linguistics.
- Ayman Alhelbawy, Mark Lattimer, Udo Kruschwitz, Chris Fox, and Massimo Poesio. 2020. An nlp-powered human rights monitoring platform. *Expert Systems with Applications*, 153:113365.
- S. S. Alizadeh and P. Moshashaei. 2015. The bowtie method in safety management system: A literature review. *Scientific Journal of Review*, 4(9):133–138.
- Patricia Bromley, Evan Schofer, and Wesley Longhofer. 2020. Contentions over world culture: The rise of legal restrictions on foreign funding to ngos, 1994–2015. *Social Forces*, 99(1):281–304.
- Thomas Carothers and Saskia Brechenmacher. 2014. Closing space: Democracy and human rights support under fire. Technical report, Carnegie Endowment for International Peace.
- Ilias Chalkidis. 2020–2022. Legal Crawler: A collection of scripts to crawl english legal corpora from open public domains. <https://github.com/iliaschalkidis/LegalCrawler/>.
- CIVICUS. 2025. People power under attack 2025: Global findings. https://monitor.civicus.org/globalfindings_2025/. Accessed July 2026.
- Anthony J DeMattee. 2019. Toward a coherent framework: A typology and conceptualization of cso regulatory regimes. In *Nonprofit Policy Forum*, volume 9, page 20180011. De Gruyter.
- Kendra Dupuy, James Ron, and Aseem Prakash. 2016. [Hands off my regime! governments’ restrictions on foreign aid to non-governmental organizations in poor and middle-income countries](#). *World Development*, 84:299–311.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- GRF CPAs & Advisors. 2019. The bow tie method addresses risk holistically. <https://www.grfcpa.com/2019/12/17/the-bow-tie-method-addresses-risk-holistically/>. Accessed July 2026.
- Erkan Gunes and Christoffer Koch Florczak. 2023. Multiclass classification of policy documents with large language models. *arXiv preprint arXiv:2310.08167*.
- Shabnam Hassani, Mehrdad Sabetzadeh, and Daniel Amyot. 2025. An empirical study on llm-based classification of requirements-related provisions in food-safety regulations. *Empirical Software Engineering*, 30(3):72.
- Andrej Karpathy. 2026. llm-wiki. <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>. GitHub Gist. Accessed July 2026.
- Yucheng Lu. 2025. [Tracking green industrial policies with LLMs: A demonstration](#). In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 1–10, Vienna, Austria. Association for Computational Linguistics.
- SearXNG Team. 2026. SearXNG: A privacy-respecting, hackable metasearch engine. <https://docs.searxng.org/>. Accessed September 2026.
- Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.
- Amy Wares. 2021. [Getting started with enterprise risk management: A guide for nonprofits](#). Technical report, ERM Initiative, Poole College of Management, North Carolina State University. Prepared with support from GRF CPAs & Advisors.
- Marianne Wilson, David M. Howcroft, Ioannis Konstantas, Dimitra Gkatzia, and Gavin Abercrombie. 2025. [Participatory design for positive impact: Behind the scenes of three NLP projects](#). In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 252–263, Vienna, Austria. Association for Computational Linguistics.
- Wolters Kluwer. The history of bowtie. Barrier Based Risk Management Knowledge Base, <https://www.wolterskluwer.com/en/solutions/enablon/bowtie/expert-insights/barrier-based-risk-management-knowledge-base/the-history-of-bowtie>. Accessed July 2026.
- Hao Zhang, Daniel Dao, James Bowden, and Mark Cummins. 2025. Large language model application for regulatory horizon scanning: Case study on esg greenwashing regulations.

A Experimental Configuration and Artefact Availability

Model and implementation. The pilot was implemented in Python using the Anthropic Messages API, the claude-sonnet-5 model, the Anthropic Python SDK, and pandas. Sampling parameters, including temperature, top_p, and top_k, were not set explicitly and therefore followed the provider

defaults. Collection, filing, and article maintenance used separate prompts and produced timestamped JSON artefacts.

Evidence collection. The collector used a trusted-source registry containing `icnl.org` and `monitor.civicus.org`, supplemented by open-web search. A single retrospective collection covered three assessment periods: the baseline at 31 December 2024, developments during 2025, and developments through 30 June 2026.

Each evidence record contained a title, factual description, event date, publication date where available, source URL, supporting source text, and document type. Stable evidence identifiers were derived from the source and record content and were used for citation and deduplication. The collector retrieved candidate evidence but did not assign threat categories, polarity, or risk status.

Evidence filing. The filing stage operated without web access and received only the records returned by the collector. For each record, the model assessed its grounding and relevance and returned zero or more assignments. Each assignment contained a threat category, subcategory, taxonomy fit, polarity, affected barrier, supporting text, rationale, and confidence score.

Polarity was classified as *restrictive*, *permissive*, or *contextual*. Permissive evidence required an affirmative and source-supported protection, repeal, withdrawal, reversal, suspension, or finding of non-applicability. Missing evidence, or the absence of a restrictive statement, was not treated as permissive evidence. Records were flagged for human review when grounding was incomplete, taxonomy fit was weak, interpretation was ambiguous, or model confidence fell below the predefined review threshold. Automated validation checked that every evidence record received a structurally valid filing result.

Sequential article maintenance. The maintenance stage also operated without web access. Article revisions were produced for 31 December 2024, 31 December 2025, and 30 June 2026. At each revision, the model received the previous version of the five threat-category articles, the newly available classified evidence, and a mechanically generated summary of cumulative evidence coverage.

All articles began as *unassessed*. The editorial policy allowed four statuses: *unassessed*, *low risk*, *high risk*, and *materialised*. Unassessed denoted absent or insufficient evidence. Low risk required

affirmative permissive evidence. High risk required supported restrictive evidence, while materialised additionally required evidence that the restriction or operational loss was currently in effect. Contextual evidence alone could not support an assessed status.

Every factual claim and status justification had to cite evidence from the corresponding threat category. Automated validation rejected unsupported or cross-category citations and status assignments that did not meet the editorial policy. The resulting assessment remained a proposal for human review rather than an autonomous organisational decision.

Run protocol. For the reported experiment, the collected evidence was frozen before the filing and maintenance stages. This held the evidence set constant while the filing instructions and four-state maintenance policy were applied. No application-level selection among alternative valid model outputs was performed.

The three-period bow-tie visualisation was generated directly from the validated article revisions and their cited evidence bases. It therefore represents the state of the maintained knowledge base rather than an independent assessment.

Artefact availability. The supplementary materials supporting the proof of concept are available in the following GitHub repository:

https://github.com/bernatcp/NGO_Bowtie