CCF Database: A Machine-Learning-Annotated Corpus of 283,964 Canadian Climate Articles (1978–2026)

Antoine Lemor*[1], Alizée Pillod†[2], Matthew Taylor‡[2], and Richard Nadeau§[2]

[1]Université de Sherbrooke, Sherbrooke, Québec, Canada
[2]Université de Montréal, Montréal, Québec, Canada

Abstract

News media shape how societies perceive and respond to climate change, yet fine-grained analysis remains hindered by fragmented corpora and ad hoc coding schemes. The Canadian Climate Framing (CCF) database answers both with a large-scale bilingual annotated corpus and a generalizable annotation framework. It covers 283,964 climate-related articles from 22 Canadian news outlets (1978–2026), processed into 9,908,776 two-sentence units (83.4% English, 16.6% French). Each unit carries 65 hierarchical binary annotations spanning eight main frames, actors, events, solutions, tone, geography, and urgency, produced by 128 BERT and CamemBERT classifiers trained on 4,000+ expert-coded sentences. Against a 1,000-sentence gold standard double-coded by an independent annotator (Cohen's $\kappa = 0.596$, Krippendorff's $\alpha = 0.698$, Gwet's AC1 = 0.894), the classifiers reach a pooled macro F_1 of 0.866 (mean per-category macro F_1 = 0.773), and every category carries an A/B/C reliability tier. The database—sentence-level annotations, per-article aggregations, and 10,192,740 BAAI/bge-m3 embeddings in six relational tables—is openly released on Zenodo (CC-BY 4.0) as a PostgreSQL edition and an Apache Parquet mirror, refreshed roughly every six months by the CCF Observatory.

*antoine.lemor@umontreal.ca
†alizee.pillod@umontreal.ca
‡matthew.taylor@umontreal.ca
§richard.nadeau@umontreal.ca

Background & Summary

On February 18, 1978, *The Globe and Mail* ran a story headlined "Some think it's warming up" that reported on a survey of leading climate experts. Interestingly, almost half a century ago, the article already laid out for its readers what have since become the well-established root causes of climate change:

> "The concern is that possible warming, the so-called 'greenhouse effect,' is generally associated with the increase of carbon dioxide in the atmosphere by the burning of fossil fuels."

This excerpt clearly embodies the role the media play in informing society and citizens, who often lack direct access to academic research. At the same time, the media also reflect the social and political realities of the country they cover. Indeed, while some newspapers choose to relay the voices of scientists, others choose to report the controversies, so that the media both inform and reflect political and social realities. Research shows that media shape not only what people think about but also how they think about it, with effects on both public opinion and political agendas [1, 2]. The media are accordingly not a neutral arena for public debate around climate change. Instead, they actively participate in defining and constructing the problems under discussion [3, 4]. Moreover, while digital and social media have grown in prominence, newspapers remain a foundational source of information that continues to shape public discourse and to influence other media formats [5, 6]. Studying print media therefore provides a vantage point on a traditional and still highly influential channel of climate communication. Most importantly, it allows for long-term historical comparison, including periods that predate the rise of social media.

The excerpt above was drawn from the oldest article of the *Canadian Climate Framing (CCF)* database, which we introduce in this Data Descriptor. The CCF database is a machine-learning-preprocessed corpus of 283,964 climate-related newspaper articles published between 1978 and 2026 across 22 major Canadian news outlets, accompanied by a rigorous, theoretically grounded annotation framework that we release alongside the data. The corpus contains 9,908,776 two-sentence analytical units, all annotated within a 65-category hierarchical coding scheme. Because Canadian newspaper articles are protected by copyright, the deposit does not redistribute the raw sentence text of the corpus at large (the sole, clearly bounded exception—the expert-coded training and gold-standard sentences—is documented in Data Records). What is shared instead is (i) the 65-category hierarchical annotation framework—a theoretically grounded coding scheme that integrates main frames, actors, events, solutions, emotional tone, geographic focus, and urgency into a single sentence-level taxonomy—and (ii) the full output of an intensive annotation pipeline built around that framework: 65 binary sentence-level labels per analytical unit, the named-entity extractions (persons, organizations, locations), the article-level aggregates and full metadata (newspaper, date, title, author, page number), the per-category reliability tiers, and the BAAI/bge-m3 sentence embeddings (10.19 million 1024-dimensional vectors). The breadth of the annotation scheme and the embedding layer together make the deposit fully analysable at the sentence level without requiring access to the raw text. By focusing on the Canadian press, this database constitutes a historically influential source of climate information.

Canada offers a particularly compelling case for studying climate discourse. Its federal structure, bilingual culture, and vast territory create a microcosm of the diversity—and contradictions—found globally. While Canada is highly vulnerable to the impacts of climate change [7], it remains significantly dependent on fossil-fuel extraction and other resource-based industries [8]. This creates structural tensions between environmental protection and economic interests. Some provinces, notably Quebec and British Columbia, tend to adopt more progressive climate policies, whereas others with resource-intensive economies are more cautious or resistant [9]. This combination of political decentralization, linguistic duality, and regional economic variation produces diverse policy approaches and, in turn, shapes how climate change is discussed in the media. Yet systematic analysis of climate communication in Canada and elsewhere has long been constrained by methodological obstacles: until recently, large-scale corpus collection and precise, reproducible content analysis were virtually impossible. Supplementary Table S1 provides an overview of prior studies in this area, highlighting their scope, focus, and coverage. The dominant pattern is one of corpora restricted to a few outlets, short observation windows, and English-language sources.

The CCF Database was designed to address four major limitations in previous research on climate change media coverage in Canada. First, the CCF draws on a larger set of news outlets and articles, comprising 283,964 articles from 22 news outlets with substantial representation for each outlet. Previous studies generally relied on smaller corpora and a limited number of newspapers, thereby capturing a narrower media landscape. Second, the CCF spans nearly five decades, providing a broader temporal perspective than most existing studies, which often focus on short periods or extend only from the late 1980s to the 2010s. Third, while climate attitudes vary significantly across regions [10], regional climate coverage remains understudied. Existing studies generally focus on nationally circulating newspapers—most commonly *The Globe and Mail*, *National Post*, and *Toronto Star*—and rarely compare a large set of provinces and territories. The CCF therefore includes a wide range of regional newspapers. Finally, the CCF reflects Canada's linguistic diversity through a bilingual corpus that includes a substantial share of French-language content (83.4% English, 16.6% French), with francophone outlets both inside and outside Quebec. While Canada is officially bilingual, previous research has largely excluded French-language media from analysis. Two studies, in particular, have examined francophone coverage: Young and Dugas [11], comparing climate coverage between French-language outlets from Quebec and English-language outlets from the rest of Canada, and King et al. [12], analysing climate change's public-health framing with partial inclusion of Quebec.

Methodologically, the CCF database is the first climate media corpus to offer sentence-level annotations derived from supervised machine learning and natural language processing (NLP). Unlike unsupervised approaches such as topic modelling, which remain dominant in the field, the CCF methodology relies on transformer-based classifiers (BERT for English, CamemBERT for French) trained on over 4,000 manually expert-coded sentences. NLP enables scholars to examine much larger corpora in far less time, opening new avenues for systematic inquiry [13, 14, 15]. However, as summarised in Supplementary Table S1, when climate researchers do employ NLP it is most often to identify key actors through named-entity recognition, to detect stance, or to conduct sentiment analysis, which tend only to map actors, tone, or evaluative direction rather than interpretive structures. The dominant computational approach in the field is still unsupervised topic modelling, where topics

are treated as proxies for frames—an assumption that Nicholls and Culpepper [16] and Otmakhova, Khanehzar, and Frermann [17] flag as problematic, as topics are too narrow to capture the defining features of frames, which, following Entman [18], require identifying a problem, specifying causal interpretations, moral evaluations, and possible solutions. By contrast, the supervised approach adopted in the CCF Database enables the identification of frames, actor types, events, solution strategies, and emotional tone with a level of semantic precision that unsupervised methods cannot achieve. Only a few studies have applied comparable transformer-based models in this domain: Luo, Card, and Jurafsky [19] trained a single BERT classifier to analyse the evolution of climate partisanship in U.S. media; Meddeb et al. [20] used a CamemBERT model to detect climate misinformation in French news; Frermann et al. [21] fine-tuned transformer models to detect narrative roles (conflict, villain, resolution) in media articles on three issues including climate change; Badullovich et al. [22] benchmarked a RoBERTa classifier against zero-shot LLMs on issue and rhetorical frames in climate change editorials published in *Nature* and *Science*. None of these works, however, have trained and deployed a comparable suite of transformer models to systematically annotate climate-related media coverage at the scale, breadth, and bilingual coverage of the CCF database.

Validated against an independent gold standard of 1,000 expert-coded sentences whose reliability was confirmed through double-coding (Cohen's $\kappa = 0.596$, Krippendorff's $\alpha = 0.698$, Gwet's AC1 $= 0.894$, on the blind-coding phase of sentences 601–1000), the CCF trained models attain a pooled macro F1 score of 0.866 across the 65 annotation categories (0.773 when the 65 per-category macro F1 scores are averaged without pooling; both quantities are defined and discussed in the Technical Validation section). In addition, named-entity recognition (NER) was applied to extract persons, organisations, and locations mentioned in the articles, and the entire corpus was embedded with the BAAI/bge-m3 multilingual encoder so that semantic-similarity queries can be issued directly against the deposited database. With 128 BERT and CamemBERT classifiers covering frames, actors, events, solutions, and emotional tone, the CCF database fills a clear methodological gap by introducing a supervised NLP approach capable of detecting multiple analytical dimensions at scale, and the absence of raw text in the deposit is offset by the depth of the annotation: every sentence carries 65 binary labels, a deduplicated NER signature, and a 1024-dimensional embedding, so researchers can run frame-based, actor-based, and semantic queries directly on the deposited tables. This technical paper presents the methodology underlying the CCF database and its structure. The remainder of the document outlines the full workflow (data acquisition and preprocessing, model training, validation, and database enrichment) and offers the conceptual rationale and practical steps necessary for researchers wishing to work with the CCF database or to adapt its methodology elsewhere.

Methods

The CCF Database was built through a mixed-methods design that coupled large-scale data collection with state-of-the-art machine learning (ML). The resulting pipeline provides a systematic and reproducible framework for large-scale analysis of climate discourse that allows researchers to track how climate change is discussed, how narratives evolve over time, and

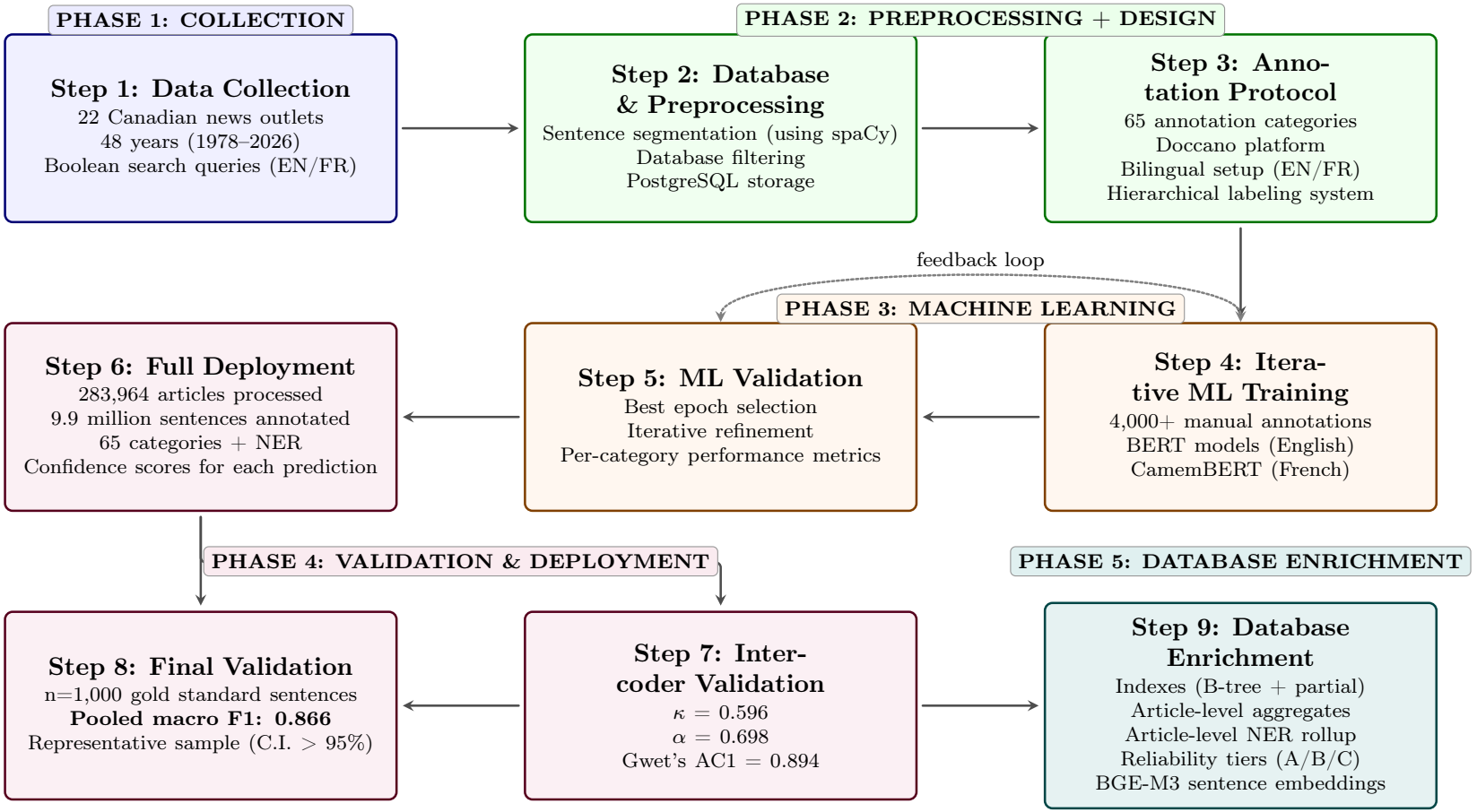

Figure 1: Methodological pipeline of the Canadian Climate Framing (CCF) project.

who participates in the conversation.

Figure 1 presents the complete methodological pipeline and illustrates how the CCF Database was assembled from raw newspaper articles into a fully annotated database suitable for robust and reproducible climate communication research, both quantitative and qualitative. The pipeline combined human expertise with computational methods through an iterative human-in-the-loop machine learning procedure. This approach ensured both the scalability needed to process 283,964 climate-related articles and the accuracy required for rigorous academic research.

The methodology comprises five distinct phases: (1) the *Collection Phase* involved systematic gathering of climate-related articles from 22 major Canadian news outlets using search queries resulting in 283,964 articles spanning 1978–2026; (2) the *Preprocessing and Design* phase transformed raw text into structured data, including sentence segmentation, language verification, and development of a comprehensive 65-category annotation framework; (3) the *Machine Learning* phase created and validated transformer-based machine learning models (BERT for English, CamemBERT for French) through iterative training with over 4,000 expert manual annotations; (4) the *Validation and Deployment* phase ensured rigorous quality assurance by combining standard performance metrics with a human evaluation protocol based on an expert-coded gold standard that was independently double-coded by a second annotator to assess intercoder reliability, before deploying the trained models to annotate the complete corpus of 9.9 million sentences; and (5) the *Database Enrichment* phase consolidated the sentence-level output into a publication-ready relational system by adding four derived tables (article-level annotation aggregates, article-level named-entity rollups, per-category reliability tiers, and 10.19 million BAAI/bge-m3 sentence-and-title embeddings indexed for approximate nearest-neighbour search), together with a complete index structure, so that the deposit can be queried directly without rewriting the same aggregation logic on every project.

Phase 1 & 2: Data collection, preprocessing, database creation and annotation protocol

Step 1: Data collection

The CCF database was built through a systematic protocol to trace the evolution of climate discourse in Canada's bilingual print media, reaching back to the earliest available archives.

Articles were retrieved through university library access to three press databases: Factiva, Eureka.cc, and ProQuest Canadian Major Dailies. These platforms are the standard providers of full-text archives for Canadian newspapers and are routinely used in peer-reviewed media research [11, 12, 20]. Each platform was queried independently with the same bilingual Boolean keywords (see Table 1 below) restricted to the selected outlets and to the full available time range (the 20 newspapers of the original collection; the CCF Observatory subsequently added the two public-broadcaster news sites through their own feeds), and duplicate articles appearing in more than one platform were collapsed at preprocessing time on the basis of newspaper, date, headline, and a 95% fuzzy similarity of the article body (see Step 2). Full bibliographic metadata are included in the deposit so that the corpus remains transparent and reproducible while staying within the bounds of copyrights.

The design relies on the fair-dealing research exception in Canadian copyright law and the analogous fair-use research exception in U.S. law, both of which authorise non-consumptive, transformative use of copyrighted text for academic research; researchers with institutional access to Factiva, Eureka.cc, or ProQuest Canadian Major Dailies can recover the original text from the bibliographic coordinates.

A total of 22 news outlets—20 newspapers and, since the corpus came under the stewardship of the CCF Observatory, the two news websites of the Canadian public broadcaster (CBC.ca and Radio-Canada.ca, covered from 2023 onward)—are included in the corpus (see Supplementary Table S2). The objective was to include both national and regional outlets in order to ensure broad geographic representation. Within this pool, newspapers were primarily chosen based on circulation levels [23], with the highest-circulating outlets retained. Additional criteria also informed the selection, including the desire to capture a range of political orientations and to build a bilingual corpus that reflects Canada's linguistic landscape. The corpus accordingly includes French-language outlets both inside and outside Quebec, together with one English-language outlet located in Quebec; this asymmetry mirrors the country's linguistic geography, in which Quebec is the only province with French as its sole official language while New Brunswick is officially bilingual and several other provinces—most notably northeastern Ontario along the Quebec border—host sizeable francophone communities. The final selection was also partly shaped by availability constraints. When the highest-circulating newspaper of a given province was unavailable, the next highest-circulating outlet was retained instead; for Prince Edward Island, Nunavut, and the Northwest Territories, no eligible newspaper was available for inclusion at all. Coverage windows are bounded by the access periods of the source aggregators: Le Droit is covered until December 2023, The Whitehorse Daily Star until May 2024, and Le Devoir and L'Acadie Nouvelle until March 2025, while all other outlets extend to the corpus cut-off of July 2026; Supplementary Table S2 details the composition and per-outlet counts.

The corpus spans 48 years, from 1978 to 2026, with 1978 marking the earliest article identified, quoted in the introduction of this paper. To identify climate-related articles, the authors used a transparent keyword-based filter with seven climate-domain terms per language (Table 1). To be included in the data set, an article had to contain at least one of these keywords in its title or body. These queries yielded an initial corpus of over 300,000 articles. As with any keyword-based retrieval, the procedure imposes a recall ceiling: it cannot capture paraphrastic, metonymic, or strictly context-only mentions of climate change, and the deposit therefore covers articles that explicitly use one of the query terms rather than the broader set of articles that discuss climate change in some form.

Table 1. Boolean queries used for article retrieval

Language	**Boolean Query**
English	"global warming" OR "climate change" OR "climate disruption" OR "climate disturbance" OR "climate disturbances" OR "climate crisis" OR "greenhouse gas"
French	"réchauffement climatique" OR "réchauffement planétaire" OR "changement climatique" OR "changements climatiques" OR "dérèglement climatique" OR "crise climatique" OR "gaz à effet de serre"

Step 2: Preprocessing

The preprocessing phase transformed the raw data into a structured dataset suitable for machine learning. Each article was segmented into two-sentence contexts using language-specific spaCy models—`en_core_web_lg` for English text and `fr_dep_news_trf` for French text, both optimised for news content. Sliding windows with single-sentence overlaps were implemented to ensure complete coverage while maintaining semantic continuity throughout each article. This level of granularity supported machine-learning applications by increasing the size of the training dataset. More importantly, working at the (near-)sentence level was the only approach that allowed ex-post analysis to remain highly granular compared to paragraphs, where multiple frames, actors, or solutions can co-occur and become indistinguishable. Article-level variables could then be reconstructed with high granularity by aggregating sentence-level annotations (e.g., percentage of sentences in an article that mention a specific frame or actor), whereas paragraph-level annotations would have been less precise and more difficult to interpret.

Quality control measures were applied systematically throughout the data preparation pipeline. Duplicate articles were identified and removed with fuzzy string-similarity algorithms set to a 95% threshold. Any article containing fewer than 100 words was also removed to eliminate reader letters, which often did not provide enough material to identify the presence of a frame and did not qualify as a press article per se. However, all other types of text (opinion columns, letters to the editor, news articles, editorials, etc.) were retained. Regarding this matter, Lawrence [24] explains that it is precisely in editorials, or more generally in opinion texts, where journalistic standards are applied with a greater degree of flexibility, that the battle of framing is best observed. Articles published by external news agencies (for example, *Agence France-Presse*, *CanWest News Service*, *Reuters*, or *Bloomberg*) or by other newspapers not included in the analysis but which often belong to the same media conglomerate as the selected outlets (for example, *the Financial Post*) were also retained. These republications were treated as aligned with the editorial direction of the host news outlet, and a republication did not contradict the objectives of the study since it indicated editorial interest in a topic concerning climate change. This process reduced the corpus to 283,964 unique articles. Duplicates of the same article appearing in different media outlets were deliberately kept, since their co-occurrence is itself an object of analysis when measuring media bubbles.

Language was then verified with the pre-trained fastText language-identification model [25, 26], in order to ensure consistent categorisation of bilingual content, and publication dates were standardised at the same step. This data preparation produced a clean, standardised corpus ready for machine learning and annotation, stored in two PostgreSQL tables: one with full metadata (titles, main text, authors, dates, and page numbers) and another with the processed sentence segments and their corresponding identifiers.

The resulting database is illustrated in Figure 2, which displays the distribution of article counts by media (22 outlets), while Figure 3 shows the total article counts per year across the full time span; the final months of 2026 are incomplete by construction, the corpus being cut off at 31 July 2026. The corpus is now maintained by the CCF Observatory, the continuous extraction–annotation pipeline described above, and is refreshed on Zenodo approximately every six months. Finally, Figure 4 presents the distribution of articles by province, based on

the primary location of each outlet. This geographic breakdown shows the representativeness of the database across Canada, with 83.4% of sentence units in English and 16.6% in French. The national outlets—*The Globe and Mail*, the *National Post*, the *Toronto Star*, and the two public-broadcaster news sites—are omitted from the provincial breakdown because of their pan-Canadian scope; nevertheless, they account for 37.5% of all articles.

Step 3: Annotation protocol

The annotation framework constitutes the conceptual core of the CCF database. It comprises 65 categories organised into a hierarchical taxonomy designed to capture multiple dimensions of climate discourse at near-sentence granularity. The framework was developed through a combined theory-driven and data-driven procedure: it drew on established scholarship in framing, agenda-setting, climate communication, and social problems studies, and was refined inductively through iterative engagement with the corpus and extensive pretesting. The full taxonomy and operational definitions are reported in Supplementary Table S3, while Table 2 provides a compact overview of the framework's structure. The framework was organised around *Primary Categories* corresponding to broad analytical dimensions, and *Sub-categories* capturing internal distinctions within each *Primary Category* (Table 2).

Primary categories. The framework was organised around four primary analytic categories that structure the annotation protocol: *Main Frames*, *Actors/Messengers*, *Events*, and *Solutions* (Table 2). Drawing on climate communication research, these categories acknowledge that media coverage does more than describe the issue, since it also selects the interpretive lens through which it is read and understood, the actors who are granted voice, the events that anchor attention, and the courses of action that are put forward as plausible.

First, for *Main Frames*, the authors built on the general framing tradition often associated with Entman [18], which encompasses four dimensions: problem definition, causal interpretation, moral evaluation, and treatment recommendations. In practice, however, applying this framework to climate discourse presents significant challenges. Individual statements frequently span multiple dimensions, and "treatments" (i.e., proposed policy responses) proved particularly difficult to identify consistently [27]. This difficulty was amplified by the nature of climate change as a "wicked problem," where proposed solutions remain subject to scientific uncertainty and political contestation [28]. For these reasons, many studies do not apply Entman's four dimensions strictly [29]. The CCF framework addressed these challenges by operationalising main frames around major societal domains in which climate change is repeatedly interpreted in media discourse: economic, security, public health, cultural, environmental, political, scientific, and justice spheres (Table 2 and Supplementary Table S3). This domain-based approach retained the core elements of problem definition, causal interpretation, moral evaluation, and treatment recommendations where they naturally emerged in the data.

Second, the framework treated *Solutions* as a distinct primary family rather than embedding them inside each main frame. This choice reflected both the empirical ambiguity of policy prescriptions in climate coverage and the substantive interest in distinguishing mitigation from adaptation, the latter having historically received less media attention [30]. Third, the *Actors/Messengers* family was grounded in the long-standing finding that *who* speaks about climate change matters for credibility, salience, and persuasion [31, 32]. Finally, the

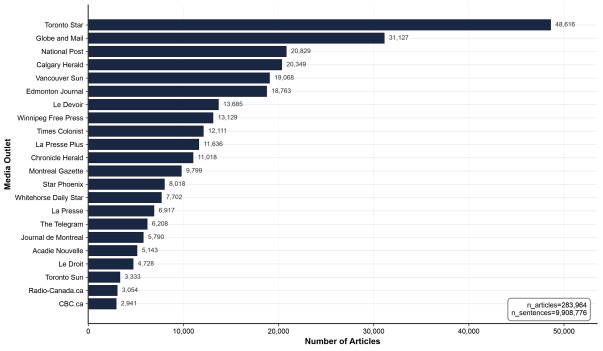

Figure 2: Distribution of articles by media.

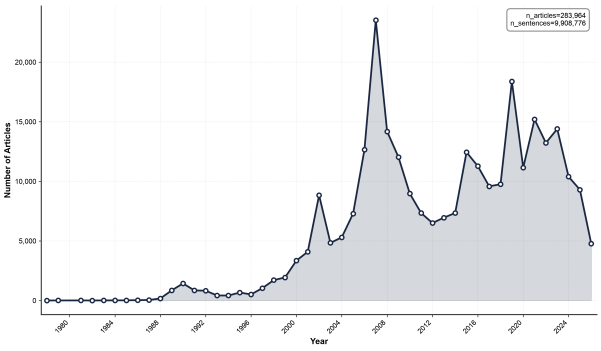

Figure 3: Total number of articles per year.

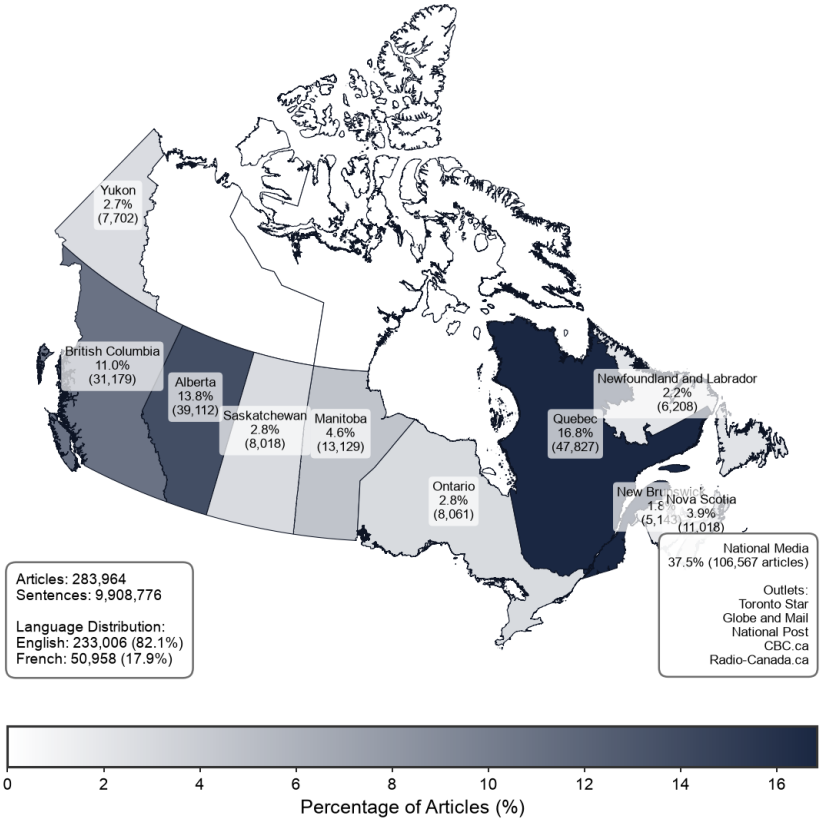

Figure 4: Distribution of articles by province.

Events family reflected agenda-setting scholarship showing that media attention is cyclical and often driven by external shocks and institutional moments (e.g., disasters, major conferences, reports) [33, 34]. Together, these four primary categories provided a theory-informed structure for capturing how climate change is framed, who is authorised to speak, what triggers attention, and which actions are articulated (Table 2 and Supplementary Table S3).

Operationalisation: hierarchical taxonomy and sub-categories. Operationally, the framework implemented a hierarchical labelling logic that mirrors its conceptual structure (Table 2 and Supplementary Table S3). Each *Primary Category* contains (i) a detection-

Table 2: The CCF annotation framework: 65 hierarchical categories for comprehensive climate discourse analysis

Main Category	N[a]	Main category detection *(in italics)* and sub-categories
Primary Categories: Main Frames (8 frames and their sub-categories)		
Economic	6	*Economic frame*, negative/positive impacts on the economy, economic disadvantages/benefits of climate action, carbon footprint of the economic sector
Health	3	*Health frame*, negative impacts on health, health co-benefits of climate action
Security	5	*Security frame*, climate refugees, conflict, post-disaster military assistance, disruption of military operations
Justice	6	*Justice frame*, winners/losers of climate action, differentiated responsibility, unequal vulnerability, unequal access to climate action, intergenerational justice
Political	5	*Political frame*, policy action, political debate, political positioning, public opinion data
Scientific	5	*Scientific frame*, scientific debate, popularisation or discovery, questioning of climate science, defense of climate science
Environmental	3	*Environmental frame*, loss of natural environments, loss of fauna and flora
Cultural	5	*Cultural frame*, artistic representation, disruption of cultural or sports events, loss of indigenous practices, carbon footprint of the cultural and sports sectors
Primary Categories: Actors, Events and Solutions (and their sub-categories)		
Actors/Messengers	9	*Any messenger*, scientists, politicians, activists, medical/economic/security/legal/cultural experts
Events	6	*Any event*, natural disasters, conferences, reports, elections, policy announcements
Solutions	3	*Any solution*, mitigation, adaptation
Emotional Tone	3	Positive (hope, optimism), negative (fear, anxiety), neutral/none
Geographic Focus	1	Canadian places, actors, data, and policies
Urgency to act	1	Conveys immediate danger, crisis, or "code red" urgency
Named Entities	3	Persons (PER), organizations (ORG), locations (LOC)
Total: 65 categories		

[a] N counts the main category detection (shown in italics) together with its sub-categories; e.g., the Economic frame: 1 main category detection + 5 sub-categories = 6. Flat categories (Emotional Tone, Named Entities) and single-detection categories (Geographic Focus, Urgency to act) have no separate main category detection.

level category that captures whether the broader dimension is present in a two-sentence unit, and (ii) fine-grained sub-categories that specify which internal distinctions apply when the primary category is detected. This design enabled multi-label annotation: multiple frames, actors, events, and solutions could co-occur within the same unit, and the same applied to sub-categories. Within *Main Frames*, each domain frame (e.g., Economic, Health, Justice) was paired with sub-categories capturing recurring interpretive distinctions observed in the literature and in the corpus (Table 2 and Supplementary Table S3). Importantly, these sub-categories were not treated as mutually exclusive: a unit could simultaneously describe, for instance, the negative economic impacts of climate change and the economic benefits of climate action. The same principle applied to *Actors/Messengers*, where a single person could legitimately be coded as multiple expert types (e.g., a physician would be classified as having both medical and natural science expertise). For full category definitions, see Supplementary Table S3.

Other dimensions: context, tone, urgency, and entity extraction. In addition to the four *Primary Categories*, the framework included theoretically motivated dimensions that capture context and meaning-making beyond framing in the strict sense (Table 2 and Supplementary Table S3 and S11). A *Geographic Focus* category recorded whether coverage situated climate change in a Canadian context, reflecting concerns that climate change is often portrayed as geographically distant [35]. An *Emotional Tone* category captured positive, negative, or neutral valence, consistent with research showing that emotions shape how audiences interpret and respond to climate discourse [36, 37]. An *Urgency to act* category captured alarmist or "code red" signals that communicate immediacy and crisis framing. Finally, Named Entity Recognition (NER) extracted and classified *Persons*, *Organizations*, and *Locations* into a JSON field, enabling downstream analyses such as cocitation networks and the mapping of epistemic authorities (Table 2; Supplementary Table S3).

Pretesting and iterative refinement. Finally, alongside these theoretical design choices, the framework was iteratively refined through inductive coding rounds. As sentences were annotated, emergent themes sometimes prompted targeted additions and clarifications to category definitions, followed by pretesting to ensure internal consistency and coder interpretability. This procedure strengthened the framework's construct validity and ensured that the final taxonomy remained both theoretically anchored and empirically adequate. The following real example from the database—an excerpt from a 2022 article—illustrates the corresponding manual annotation protocol:

Example (annotated sentence).

"In Scientific American, professor of Environmental Studies at Humboldt State University Sarah Jaquette Ray writes, 'Climate anxiety can operate like white fragility, sucking up all the oxygen in the room and devoting resources toward appeasing the dominant group.' But it would be a mistake to conclude that this term isn't applicable in the Global South, period."

Tagged Dimensions	**Tagged categories**
Actors/Messengers	Scientists
Main frames	Justice (primary category) + North–South responsibility; Health (primary category)
Emotional tone	Negative

Phase 3: Machine learning

The development of the training data followed an iterative manual annotation protocol aligned with the annotation framework previously described. A single expert annotator (climate policy and communication) labelled sentences in sequential batches (first 1,000 sentences, then three additional batches of 1,000 each) for a total of 4,000 annotated samples. The choice of a single highly-trained expert during training, with an independent second coder reserved for the gold-standard validation, was made to maximise definitional consistency across the 4,000 training sentences. In framing analysis, where 65 conceptually overlapping categories must be applied to fine-grained sentence units, multi-coder training can introduce drift in operational definitions that exceeds the noise it removes; the more reliable protocol is therefore to stabilise the codebook with a single expert and to validate it against an independent coder on a held-out gold standard [38]. The batching served two explicit objectives that were only attainable with a single coder: (1) to iteratively improve and stabilise the annotation guidelines after each round; and (2) to progressively monitor and optimise model training to reach the highest attainable F1 score. This process culminated in a training-phase macro F1 score of 0.826—the unweighted mean of the per-category macro F1 scores obtained on each category's held-out validation split (see Table 3, and Supplementary Tables S4 and S11 for detailed metrics).

The sampling process drew equally from English and French language groups, with final annotation counts reaching 1,927 English and 2,073 French sentences. The training dataset was subsequently partitioned using stratified random sampling to create training (80%) and validation (20%) databases for each category and language. Supplementary Table S5 provides the complete distribution of training and validation samples for all annotation categories. The random sampling approach ensured broad coverage across the temporal span and diverse media sources. The annotation protocol was developed with detailed operational definitions for each of the 65 categories to ensure consistency.

The machine learning pipeline leveraged transformer architectures optimised for each language through a custom fork of the AugmentedSocialScientist library [39, 40]. The fork, built on the original library of Do, Ollion, and Shen [39], added several functionalities that were central to ensuring robust machine-learning training (metric logging at every training epoch, best-model selection using a weighted F1 score ($0.98 \times F_1^{(1)} + 0.02 \times$ macro-F_1), and an automated reinforced training protocol for underperforming models, described below) [40]. English text processing employed BERT-base-cased, while French text used CamemBERT-base, both containing 110 million parameters. The English models being case-sensitive, their tokeniser must be instantiated with `do_lower_case=False`; a `tokenizer_config.json` enforcing this ships with every released English model directory. The training strategy implemented hierarchical classification, where detection models were trained first as binary classifiers on the full annotated dataset, followed by sub-classification models trained exclusively on manually annotated sentences that were positive for the corresponding primary category. This approach minimised false-positive propagation and allowed for specialised optimisation for each classification task.

All 128 models shared a single training configuration, reported here in full for reproducibility. Optimisation used AdamW ($\epsilon = 10^{-8}$) with a fixed random seed of 42 (Python, NumPy, and PyTorch generators) for shuffling and weight initialisation; the reinforced-phase sampler

draws from PyTorch's default global generator, a known limitation recorded in the deposited configuration file. The training loss was a standard, unweighted two-class cross-entropy in the normal phase. The normal training phase ran for up to 20 epochs with a batch size of 32 and a learning rate of 5×10^{-5}; the reinforced phase (described below) ran for up to 20 additional epochs with a batch size of 64 and a learning rate of 5×10^{-6}. Input sequences were tokenised to a maximum length of 512 tokens. At every epoch the pipeline logged precision, recall, and F1 on the held-out validation split, and the best epoch was selected by the weighted criterion $0.98 \cdot F_1^{(1)} + 0.02 \cdot \text{macro-}F_1$ introduced above. All training ran on a single Apple Mac Studio (M2 Ultra, 128 GB unified memory); the exact static configuration and the per-model hyperparameters are archived in the deposit (`Database/Training_data/training_static_configuration.csv` and `training_hyperparameters_normalized.csv`) and summarised in Supplementary Table S11.

Models failing to achieve a positive-class F1 score of at least 0.70 during the training phase underwent an automated reinforced training protocol. This protocol was triggered for 45 of the 128 trained models (35%), concentrated in low-prevalence sub-frames—cultural, justice, and security—and in the fine-grained political and scientific sub-categories. The reinforced phase implemented weighted random sampling to oversample minority classes, doubled the batch size to 64, reduced the learning rate to 5×10^{-6}, and applied a class-weighted cross-entropy loss with a positive-class weight of 3.5. Training was extended for an additional 20 epochs, after which the best epochs of the normal and reinforced phases were compared on

Table 3: Model training performance metrics for primary annotation categories

Category	**F1 (Class 1)**		**F1 (Class 0)**		**Macro F1**	
	EN	**FR**	**EN**	**FR**	**EN**	**FR**
Main Frames						
Economic Frame	0.745	0.814	0.944	0.957	0.845	0.885
Health Frame	0.800	0.667	0.989	0.994	0.894	0.830
Security Frame	0.870	0.800	0.996	0.997	0.933	0.898
Justice Frame	0.719	0.717	0.975	0.981	0.847	0.849
Political Frame	0.808	0.774	0.897	0.888	0.853	0.831
Scientific Frame	0.784	0.702	0.953	0.962	0.869	0.832
Environmental Frame	0.842	0.625	0.989	0.980	0.915	0.802
Cultural Frame	0.773	0.833	0.986	0.993	0.879	0.913
Other Primary Categories: Actors, Events and Solutions						
Presence of Messengers	0.912	0.929	0.904	0.915	0.908	0.922
Presence of Events	0.794	0.819	0.935	0.932	0.865	0.876
Presence of Solutions	0.737	0.878	0.914	0.944	0.825	0.911
Mention of Canada	0.942	0.964	0.968	0.980	0.955	0.972
Urgency to act	0.591	0.649	0.975	0.984	0.783	0.816
Overall Average*	**0.794**	**0.782**	**0.956**	**0.962**	**0.875**	**0.872**
Combined (ALL)	**0.753**		**0.900**		**0.826**	

*Overall average across all active annotation categories (see Supplementary Table S4 for complete metrics).

a combined criterion that gave 98% weight to the positive-class F1 and 2% to the macro F1, i.e. $0.98 \cdot F_1^{(1)} + 0.02 \cdot \text{macro-}F_1$. The reinforced epoch was retained as the final model for 23 of these 45 cases, while the remaining 22 reverted to their best normal-training epoch; averaged over the 23 reinforced models eventually retained, the positive-class F1 improved by +0.25, lifting most of the affected categories above the deployment threshold. Two categories—*positive impacts of climate change on health* and *carbon footprint of the health sector*—could not be trained in either language because the manually annotated sentences contained no positive examples for them; they were consequently excluded from the deployed framework, leaving 65 deployed categories out of the 67 candidate categories listed in Supplementary Table S3. Two further categories—*post-disaster military assistance* and *disruption of military operations due to climate change*—could not be trained in English for the same reason but were successfully trained in French. These two remain part of the 65 deployed categories: they are annotated on French sentences only, their English annotations are left as `NULL` by construction (and coalesced to zero in the article-level aggregates), and both are flagged Tier C in the reliability lookup (Supplementary Table S10). The deployed suite therefore counts $65 \times 2 - 2 = 128$ models. The complete per-model training provenance, including which models triggered reinforcement and which retained the reinforced epoch, is provided in Supplementary Table S11.

Phase 4: Validation and deployment

Phase 4 chained the full deployment of the trained models (Step 6) with the two validation steps (Steps 7 and 8) that closed the methodological pipeline. Step 6 is described immediately below; the design of the validation steps (sampling protocol, gold-standard construction, model performance metrics, and inter-coder reliability) is described separately and in full in the Technical Validation section below.

Step 6: Full deployment

Once the per-category models had been finalised, all trained models were applied to annotate the complete CCF Database of 9.9 million sentences across 283,964 articles. A hierarchical classification strategy was implemented that directly mirrored the annotation protocol's structure of primary categories and their associated sub-categories, as illustrated in Figure 5.

The hierarchical approach reflected the conceptual organisation of the annotation framework with eleven primary detection categories, each with specific sub-categories. The eight main frames (*Economic*, *Health*, *Security*, *Justice*, *Political*, *Scientific*, *Environmental*, and *Cultural Frame*) functioned as primary categories with their internal distinctions—for example, *Economic Frame*, when positive (=1), triggered five economic sub-models (*negative/positive economic impacts of climate change*, *economic disadvantages/benefits of climate action*, *carbon footprint of the economic sector*), while a negative detection (=0) bypassed these entirely.

Similarly, the three other primary categories followed this conditional logic: *Actors/Messengers Detection* activated nine actor-type sub-models (*Health Expert*, *Economic Expert*, *Security Expert*, *Legal Expert*, *Cultural Expert*, *Natural Scientist*, *Social Scientist*, *Activist*, *Public Official*), *Event Detection* triggered eight event-type classifications (*Ex-*

treme Weather Event, *Meeting/Conference*, *Publication*, *Election*, *Policy Announcement*, *Judiciary Decision*, *Cultural Event*, *Protest*), and *Solutions Detection* applied two solution sub-categories (*Mitigation Strategy*, *Adaptation Strategy*). The deployment also processed standalone primary categories without hierarchical structure—*Emotional Tone* (*Positive Emotion/Negative Emotion/Neutral Emotion*), *Geographic Focus* (*Canadian Context*), *Urgency/Alarmism*.

In addition to the custom-trained classification models, three pre-trained Named Entity Recognition (NER) models were employed, selected on the basis of empirical evaluation on the CCF data. A hybrid approach tailored to each language was implemented. For English texts, the publicly available `dslim/bert-base-NER` model (hosted on the Hugging Face Hub) was used for all three entity types: persons, organisations, and locations. For French texts, person entities were extracted with the spaCy `fr_core_news_lg` pipeline, while organisations and locations were extracted with the `Jean-Baptiste/camembert-ner` model (also hosted on the Hugging Face Hub). This hybrid strategy was chosen to leverage the respective strengths of each model: spaCy's superior performance on French person names, and CamemBERT-NER's robust identification of French organisational and geographical entities. Per-language performance metrics for the three NER models are reported in Supplementary Table S6.

Phase 5: Database enrichment

The deployment described above produced the two original tables of the CCF Database (`CCF_full_data` and `CCF_processed_data`). The database was then further enriched with four additional tables and a complete index structure so that researchers could interrogate the

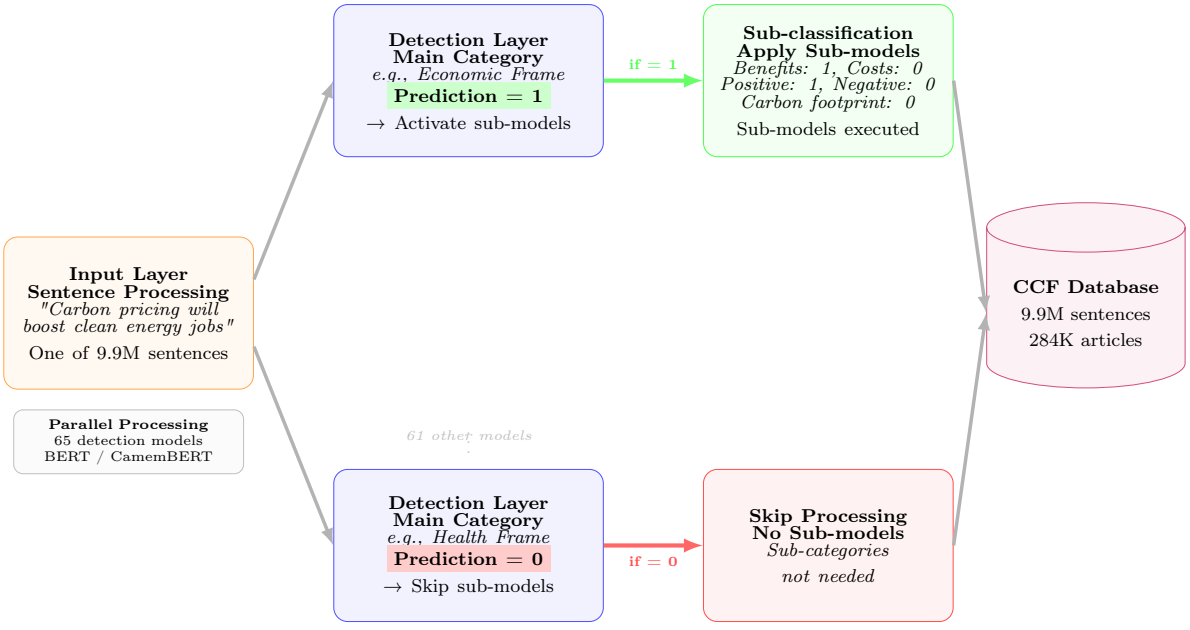

Figure 5: Hierarchical annotation pipeline. The system processes each sentence through detection models that determine whether corresponding sub-category models should be applied, enabling efficient annotation of the entire corpus.

corpus without rewriting the same article-level aggregations on every project. The enrichment procedure was deterministic and idempotent: it ran end-to-end from a freshly restored database dump, and each step (index provisioning, article-level aggregates, article-level entities, sentence embeddings, and reliability tiers) is described in detail below.

Indexes. The first enrichment step provisioned twenty indexes on the two original tables: unique primary keys on `(doc_id)` for `CCF_full_data` and `(doc_id, sentence_id)` for `CCF_processed_data`; routine B-trees on the columns researchers query most frequently (`date`, `media`, `language`); and partial B-trees on `doc_id` restricted to the rows where one of the eleven primary detection flags is positive (*economic_frame*, *health_frame*, *security_frame*, *justice_frame*, *political_frame*, *scientific_frame*, *environmental_frame*, *cultural_frame*, *messenger*, *event*, *solution*). The partial indexes accelerated the most common analytical workflow—retrieving the subset of articles or sentences in which a given frame is positive—without paying the storage cost of a full B-tree on each binary column.

Article-level aggregates. The `CCF_article_aggregates` table was built with one row per `doc_id`; for each of the 65 annotation categories, it stored the article-level proportion (`prop_X`, in $[0, 1]$) of sentences positive for category X, computed as the sum of the binary flag divided by the article's sentence count. The table also recorded the article's sentence count overall and by language (`n_sentences`, `n_sentences_en`, `n_sentences_fr`), the article's top frame (`top_frame`, the main frame with the highest proportion in the article), the proportion of that frame (`top_frame_prop`), the Shannon entropy of the eight main-frame proportions (`entropy_frames`, measuring how concentrated or diffuse the article's framing is), and the number of distinct messenger, event, and solution sub-categories present in the article (`n_unique_messenger_subs`, `n_unique_event_subs`, `n_unique_solution_subs`). Sub-categories whose parent primary detection was negative were coalesced to zero, so a `prop_X` reading of 0 unambiguously means "no sentence in this article is positive for category X", regardless of whether the sub-category was hierarchically evaluated.

Article-level named entities. The `CCF_article_entities` table consolidated the sentence-level Named Entity Recognition output (the `ner_entities` JSON column in `CCF_processed_data`) into one row per article. Three JSONB arrays (`entities_per`, `entities_org`, `entities_loc`) stored the deduplicated set of persons, organisations, and locations mentioned anywhere in the article; three counters (`n_unique_per`, `n_unique_org`, `n_unique_loc`) recorded their cardinalities; and three further counters reported how many sentences of the article contained at least one entity of each type. The presence of an entity in this rollup does not by itself disambiguate whether the entity is the speaker, the subject, or a passing reference; users wishing to recover speaker–statement attribution should combine the entity rollup with the sentence-level `messenger` family of annotations or with downstream attribution models that are out of scope for this database.

Sentence embeddings. The `CCF_sentence_embeddings` table was populated, for every sentence in `CCF_processed_data` and for the title of every article (encoded as `sentence_id` = 0), with the dense vector representation produced by the BAAI `bge-m3` multilingual encoder [41]. `bge-m3` was selected for three reasons: (i) it natively supports the 100+ languages relevant to bilingual Canadian coverage and is reported to perform competitively on French–English cross-lingual tasks; (ii) it produces 1024-dimensional dense vectors that combine acceptable retrieval quality with manageable storage overhead; and (iii) it has been stable and widely adopted since its January 2024 release, which is favourable for the long-term

reproducibility expected of a published dataset. Vectors were L2-normalised at encoding time, so cosine similarity reduces to a dot product and can be queried directly through the `pgvector` `<=>` operator. Storage used the `halfvec(1024)` type (2 bytes per dimension, identical to the upstream `float16` representation), and an HNSW index on the cosine operator class was built to deliver sub-second approximate top-k retrieval over the 9.9 million-row table.

Reliability tiers. The last enrichment step recorded, for each annotation category, a recommendation on how confidently it can be used in downstream analyses. The recommendation was summarised as a discrete *reliability tier* derived from the two evaluation criteria reported in full in the Technical Validation section below—the validation macro F_1 score against the gold standard, and the blind-phase Cohen's κ between the expert annotator and the independent second coder. The `CCF_reliability_tiers` table was built as a 65-row lookup classifying each category into one of three tiers: Tier A (*research-grade*) required $F_1 \geq 0.80$ *and* $\kappa \geq 0.60$; Tier C (*exploratory*) corresponded to $F_1 < 0.60$ *or* $\kappa < 0.40$; Tier B (*use with caution*) captured everything in between. The table also flagged the categories affected by *prevalence-induced reliability deflation*—that is, categories for which $F_1 \geq 0.70$ but $\kappa < 0.40$, where the chance-corrected agreement is depressed by very low prevalence rather than by systematic disagreement between coders [42, 43]—and the two categories without an English model (annotated on French sentences only). Because the annotation tables store one *column* per category while the tier table stores one *row* per category, the two are not combined through a relational join: `CCF_reliability_tiers` is a lookup that tells the analyst *which columns* of `CCF_processed_data` or `CCF_article_aggregates` to keep. Selecting the codes where `tier_overall = 'A'` and restricting the analysis to the corresponding columns limits it to the 27 categories that the authors recommend without reservation; a worked example is given in the Usage Notes section, and the full tier assignment by language is provided in Supplementary Table S10.

A diagram of the enriched database, together with the join paths that connect the six tables, appears in the Data Records section (Figure 6). The complete enrichment procedure ran end-to-end in approximately 2.5 hours on a single Apple Mac Studio with the M2 Ultra chip and 128 GB of unified memory.

Data Records

Before working with any table of the deposit, users are advised to consult the `CCF_reliability_tiers` lookup described in this section: it assigns each of the 65 annotation categories an A/B/C reliability tier, and the authors recommend restricting substantive conclusions to Tier A categories (27 of 65) unless the analysis explicitly accounts for the uncertainty of lower tiers. The database is versioned: new releases are published on Zenodo approximately every six months, as the CCF Observatory—the continuous extraction, annotation, and enrichment pipeline that now maintains the corpus—ingests newly published articles. The concept DOIs (10.5281/zenodo.20346363 for the PostgreSQL edition, 10.5281/zenodo.20346372 for the Parquet mirror) always resolve to the latest release, while version DOIs remain permanently resolvable so that any analysis can be reproduced against the exact release it used. This paper describes version 2.1.0.

Overview of the deposit

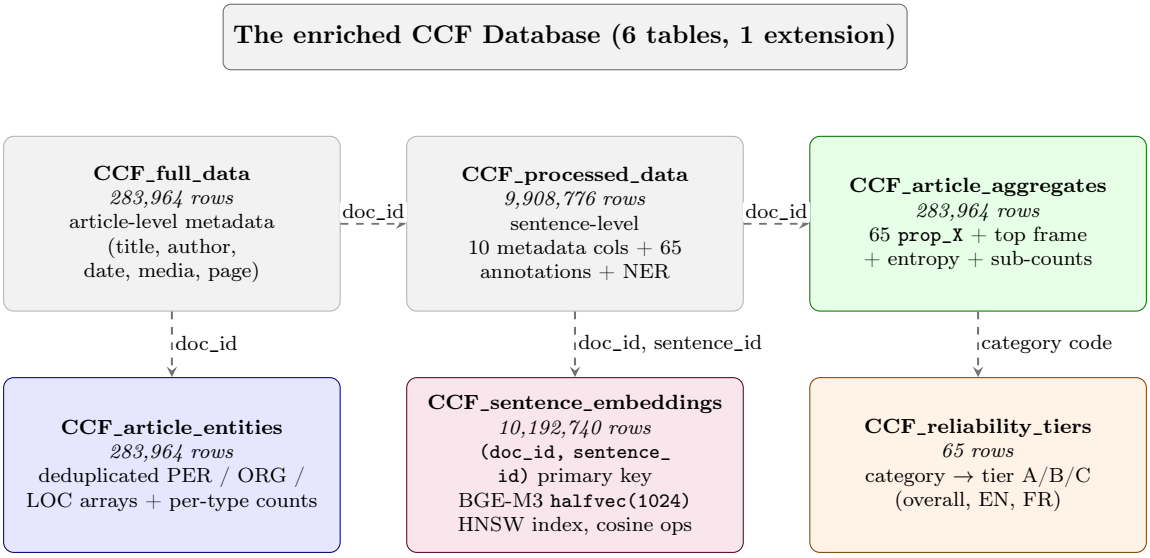

Figure 6: Schema of the enriched CCF Database. Dashed arrows indicate the join paths between tables; column types and full schemas are documented in Supplementary Table S12. `CCF_full_data` and `CCF_processed_data` are the original sentence-level resource; the three coloured groups are the article-level rollups (green), the reliability lookup (orange), and the embedding table (purple).

The CCF Database is deposited on Zenodo under a CC-BY 4.0 licence in two complementary editions sharing identical schemas. The canonical PostgreSQL edition (DOI 10.5281/zenodo.22019287 [44]) ships as a single tarball (`CCF_Database.tar`, ≈ 40 GB) wrapping a `pg_dump` *directory* archive of the six relational tables together with all B-tree, partial, and HNSW indexes; restoration requires PostgreSQL 16 or 17 with the `pgvector` extension (≥ 0.8.2 for `halfvec` support) and the deposit's README provides the exact three-line restore command (`tar -xf`, `createdb`, `pg_restore -j 8`). A sister column-oriented Apache Parquet mirror (DOI 10.5281/zenodo.22019288 [45], one ZSTD-compressed file per table, ≈ 17.5 GB) is provided for users without a PostgreSQL backend; the `halfvec(1024)` embedding column is materialised as a 1024-element `LIST<FLOAT>` and the JSONB entity arrays are serialised as UTF-8 JSON strings, so each Parquet file can be read directly with `pandas`, `polars`, R/`arrow`, DuckDB, or Spark. The companion code and documentation—the full annotation pipeline (`Scripts/Annotation/`), the enrichment and export scripts (`Scripts/Database_creation/`), and the canonical CSVs underlying the supplementary tables (`Database/Training_data/`)—are archived in a dedicated, versioned Zenodo deposit (DOI 10.5281/zenodo.22019289 [46]) and, for convenience, also bundled with both data editions as `ccf_code_and_paper.tar.gz`; a development mirror is maintained on GitHub (`antoinelemor/CCF-canadian-climate-framing`, referenced in the Code Availability statement).

Because Canadian newspaper articles are protected by copyright, the deposit does not redistribute the raw sentence text or full article bodies of the corpus at large.

What is shared is everything else: article-level metadata (titles, authors, publication dates, newspapers, page numbers), sentence-level identifiers, the 65 binary annotations, the named-entity extractions, the article-level aggregates and entity rollups, the per-category reliability tiers, and the BGE-M3 sentence embeddings. One deliberate, clearly bounded exception is made so that the single-coder training protocol and the gold standard can be independently audited: the code and documentation deposit releases the expert-coded sentences themselves—the ∼4,000 training sentences (`Database/Training_data/manual_annotations_JSONL/Annotated_sentences.jsonl`, together with the per-category train/validation splits under `annotation_bases/`) and the 1,000-sentence double-coded gold standard (`blind_verification.jsonl` and `consensus_gold_standard.jsonl`). These files contain short excerpts of two sentences each, quoted for the purpose of research and validation with full bibliographic attribution through the corpus metadata; this reproduction of brief extracts falls under the fair-dealing exception for research and private study of the Canadian *Copyright Act* (s. 29) and does not substitute for the original works. Each article in the corpus is identified by its bibliographic coordinates (`media`, `date`, `title`, `author`, `page_number`), which is the standard bibliographic reference for Canadian newspapers and is sufficient for any researcher with institutional access to Factiva, Eureka.cc, or ProQuest Canadian Major Dailies to retrieve the original text. No DOIs exist for the historical articles covered by the corpus.

The database is organised around two original tables produced by the annotation pipeline (`CCF_full_data`, `CCF_processed_data`) and four enrichment tables that pre-compute the queries most frequently issued by analysts (article-level aggregates, article-level entities, reliability tiers, sentence embeddings). The schema is summarised in Table 4 and depicted graphically in Figure 6; the join paths in that diagram point to the recommended foreign keys whenever a query needs to combine two tables.

Table 4: Overview of the six tables in the deposited CCF Database. The `pgvector` extension underpins the `CCF_sentence_embeddings` table.

Table	**Rows**	**Granularity**	**What it stores**
`CCF_full_data`	283,964	Article	Bibliographic metadata for every article in the corpus.
`CCF_processed_data`	9,908,776	Sentence	The 65 sentence-level binary annotations and the per-sentence NER output.
`CCF_article_aggregates`	283,964	Article	Per-article proportions, top frame, framing entropy and sub-category counts.
`CCF_article_entities`	283,964	Article	Deduplicated JSONB arrays of PER / ORG / LOC mentions, with per-type cardinalities and per-type sentence counts.
`CCF_reliability_tiers`	65	Category	Per-category reliability tier (A / B / C) with the underlying F_1 and κ values.
`CCF_sentence_embeddings`	10,192,740	Sentence + title	L2-normalised BGE-M3 1024-dimensional vectors, queryable through `pgvector`.

CCF_full_data – article-level metadata

This table is the canonical inventory of articles in the corpus and the natural entry point for any analysis that operates at the article level. Each of its 283,964 rows describes one unique article and is uniquely keyed by `doc_id`; every other table in the database joins back to `CCF_full_data` through this identifier. The columns are summarised in Table 5.

Table 5: Columns of `CCF_full_data`.

Column	Type	Definition
`doc_id`	`BIGINT (PK)`	Unique article identifier; foreign-key target for every other table.
`news_type`	`TEXT`	Editorial category: news, opinion, editorial, etc.
`title`	`TEXT`	Article headline.
`author`	`TEXT`	Author byline when available.
`media`	`TEXT`	Newspaper name.
`words_count`	`BIGINT`	Total word count of the article.
`date`	`DATE`	Publication date.
`language`	`TEXT`	"EN" or "FR".
`page_number`	`TEXT`	Print page (free text, e.g. "A1" or "B3").

CCF_processed_data – sentence-level annotations

This is the workhorse table of the database. It contains one row per two-sentence analytical unit, so 9,908,776 rows in total, and every annotation produced by the machine-learning pipeline lives here. The same ten metadata columns as `CCF_full_data` are denormalised on every row so that sentence-level queries can run without joining back to the article table whenever the analyst only needs the date, the newspaper, or the language. On top of that, the table stores the 65 binary annotation columns named exactly as in Supplementary

Table 6: Structural columns of `CCF_processed_data`. The 65 binary annotation columns are defined in Supplementary Table S3.

Column	Type	Definition
`doc_id`	`BIGINT`	Foreign key to `CCF_full_data`.
`sentence_id`	`BIGINT`	Position of the sentence within its article (starting at 1). `sentence_id` = 0 is reserved for the article title in the embeddings table.
(metadata)	various	Ten metadata columns duplicated from `CCF_full_data` for query convenience: `news_type`, `title`, `author`, `media`, `words_count`, `date`, `language`, `page_number`.
65 annotation columns	INTEGER (0/1, nullable)	One binary flag per category from Supplementary Table S3. Subcategories are NULL when the parent primary detection is negative.
`ner_entities`	`TEXT (JSON)`	Per-sentence NER output in the format `{"PER": [..], "ORG": [..], "LOC": [..]}`.

Table S3, and one JSON column (`ner_entities`) holding the per-sentence Named Entity Recognition output. Sub-category columns are NULL when their parent primary detection is negative; researchers who want to treat "absent" as equivalent to "not detected" should coalesce NULL to zero, which is exactly the convention adopted by `CCF_article_aggregates`. The structural columns are summarised in Table 6.

`CCF_article_aggregates` – article-level rollup

`CCF_article_aggregates` is the recommended starting point for most article-level analyses—bearing in mind the general recommendation that opens this section: its 65 `prop_*` columns inherit the reliability of the underlying sentence-level classifiers, so analyses should prioritise the columns whose category is Tier A in `CCF_reliability_tiers`. It collapses the 9.9 million sentence-level rows into one row per article (283,964 rows) and pre-computes the quantities that almost every project needs: the proportion of sentences positive for each of the 65 categories, the top frame, a measure of how concentrated or diffuse the framing of the article is, and the number of distinct sub-categories activated. Because these aggregates are computed once at enrichment time and indexed, a sentence-level `GROUP BY` on the 9.9-million-row sentence table is rarely necessary once the aggregates are available. The columns are summarised in Table 7.

Table 7: Columns of `CCF_article_aggregates`.

Column	Type	Definition
`doc_id`	`BIGINT (PK)`	Foreign key to `CCF_full_data`.
`n_sentences`	`INTEGER`	Total number of two-sentence analytical units in the article.
`n_sentences_en, n_sentences_fr`	`INTEGER`	Same count restricted to English / French sentences.
65 `prop_X` columns	DOUBLE PRECISION ($[0, 1]$)	Article-level proportion of sentences positive for category X. Sub-categories with a negative parent detection are coalesced to zero.
`top_frame`	`TEXT`	Main frame with the highest `prop_X`; NULL when no frame is positive.
`top_frame_prop`	`DOUBLE PRECISION`	Proportion of the top frame in the article.
`entropy_frames`	`DOUBLE PRECISION`	Shannon entropy (natural base) of the eight main-frame proportions; 0 when no main frame is positive.
`n_unique_messenger_subs`	`SMALLINT`	Number of distinct messenger sub-categories present in the article (0–9).
`n_unique_event_subs`	`SMALLINT`	Same for event sub-categories (0–8).
`n_unique_solution_subs`	`SMALLINT`	Same for solution sub-categories (0–2).

CCF_article_entities – article-level named-entity rollup

CCF_article_entities consolidates the sentence-level Named Entity Recognition output (the ner_entities JSON column in CCF_processed_data) into one row per article. The table is designed to answer the recurring analytical question *which entities does an article mention?* without forcing the analyst to parse the JSON for every sentence. It contains 283,964 rows and its structure is summarised in Table 8.

Table 8: Columns of CCF_article_entities.

Column	Type	Definition
doc_id	BIGINT (PK)	Foreign key to CCF_full_data.
entities_per	JSONB	Deduplicated array of persons mentioned in the article.
entities_org	JSONB	Deduplicated array of organisations.
entities_loc	JSONB	Deduplicated array of locations.
n_unique_per, n_unique_org, n_unique_loc	INTEGER	Cardinalities of the three arrays above.
n_sentences_with_per, n_sentences_with_org, n_sentences_with_loc	INTEGER	Number of sentences in the article that mention at least one entity of the corresponding type.

CCF_reliability_tiers – per-category quality lookup

CCF_reliability_tiers provides a single-row-per-category recommendation on how confidently each of the 65 annotation categories may be used in downstream analyses. The tier assignment combines validation F_1 and blind-phase Cohen's κ following the rules defined in Section ; the underlying metrics and flags are kept in the same row so that researchers can refine the tier rule for their own use case if they wish. This table is a *lookup*, not a join target: annotation categories appear as columns (not rows) in CCF_processed_data and CCF_article_aggregates, so the tier table is used to decide *which columns* to retain—select the code values where tier_overall = 'A' and restrict the analysis to those 27 columns (worked example in the Usage Notes). Columns are listed in Table 9.

CCF_sentence_embeddings – dense vector representations

CCF_sentence_embeddings enables semantic retrieval and similarity-based analyses across the whole corpus. It holds one row per sentence (and one extra row per article title, encoded with sentence_id = 0), for a total of 10,192,740 rows. Each row stores a 1024-dimensional BAAI/bge-m3 vector that is L2-normalised at encoding time, so cosine similarity reduces to a dot product and can be queried directly through the pgvector <=> operator. Storage uses the halfvec(1024) type (2 bytes per dimension, identical to the upstream float16 representation), and an HNSW index on the cosine operator class delivers sub-second approximate top-k retrieval over the full table. Columns are summarised in Table 10.

Table 9: Columns of `CCF_reliability_tiers`.

Column	Type	Definition
`number`	`SMALLINT (PK)`	Sequential index of the 65 deployed categories, in canonical framework order. Note that Supplementary Table S3 enumerates all 67 *candidate* categories (including the two excluded ones) and therefore numbers differently: categories should always be matched on `code`, never on `number`. S3 reports each category's `code` to make this crosswalk unambiguous.
`category`	`TEXT`	Readable category name.
`code`	`TEXT (UNIQUE)`	Column code as used in `CCF_processed_data`.
`tier_overall`	`CHAR(1)`	Tier on the combined corpus: "A", "B", or "C".
`tier_en`	`CHAR(1)`	Tier on the English subset.
`tier_fr`	`CHAR(1)`	Tier on the French subset.
`paradox_kappa`	`BOOLEAN`	True if the category exhibits prevalence-induced reliability deflation ($F_1 \geq 0.70$ and $\kappa < 0.40$).
`exclusion_reason`	`TEXT`	"insufficient_training_data" for the two categories that could not be trained, NULL otherwise.
`f1_macro_all, f1_macro_en, f1_macro_fr`	`DOUBLE PRECISION`	Validation macro F_1 score on the gold standard, overall and per language.
`support_all, support_en, support_fr`	`INTEGER`	Number of positive (class 1) validation examples.
`cohens_kappa_blind_phase, gwet_ac1_blind_phase, krippendorff_alpha_blind_phase`	`DOUBLE PRECISION`	Inter-coder agreement on the blind phase (sentences 601–1000).
`n_blind_phase`	`INTEGER`	Sentence count of the blind phase.
`prev_coder1_full, prev_coder2_full`	`DOUBLE PRECISION`	Prevalence (proportion of positives) per coder on the full 1,000-sentence sample.

Table 10: Columns of `CCF_sentence_embeddings`.

Column	Type	Definition
`doc_id`	`BIGINT (PK, with sentence_id)`	Foreign key to `CCF_full_data`.
`sentence_id`	`INTEGER (PK)`	Sentence position within the article. `sentence_id` = 0 is reserved for the article title.
`embedding`	`halfvec(1024)`	L2-normalised BGE-M3 vector. Cosine distance is computed via the `pgvector <=>` operator; an HNSW index on the cosine operator class is provisioned by the enrichment pipeline.

The complete data dictionary, including every column type and the full set of join paths, is provided in Supplementary Table S12.

Data Overview

This section consolidates three reproducible summaries of the deposited CCF Database. Every figure and table is generated by the script `Results/Scripts/3_data_overview.py`, which pulls the underlying numbers directly from the deposited PostgreSQL tables: no manual transcription is involved, and re-running the script regenerates the artefacts byte-for-byte.

Article-level descriptive statistics. Table 11 reports the corpus size (article count, two-sentence analytical units, and per-article word statistics) computed on `CCF_full_data` and `CCF_processed_data`, and the mean per-article share of sentences positive for each annotation category, computed on the `prop_X` columns of `CCF_article_aggregates`. The left block lists the eight frames together with the percentage of articles for which each frame is the `top_frame`; the right block lists the primary categories (messenger, event, solution, Canadian context, urgency) and the three tones.

Temporal coverage of the main frames. Figure 7 plots, for every calendar year between 1980 and 2026, the mean of each `prop_X` main-frame column of `CCF_article_aggregates`, grouped by the year extracted from `CCF_full_data.date`. No smoothing or regression is applied; each point is a yearly mean. The earliest years of the corpus carry very few articles—fewer than fifty per year before 1988 and fewer than five hundred per year before 1990—so the pre-1988 portion of the curve is computed on small samples and should be read with that caveat.

Named-entity coverage. Table 12 reports the ten most frequently mentioned entities of each NER type. Entities are extracted from the JSONB arrays `entities_per`, `entities_org`, and `entities_loc` of `CCF_article_entities`, expanded with `jsonb_array_elements_text`, filtered to remove pure punctuation and one-character tokens, and counted in articles—an entity is therefore counted once per article regardless of how many sentences mention it. Variants such as *Trudeau* vs. *Justin Trudeau* are kept separate because the underlying NER pipeline emits them as distinct strings; analysts who need canonicalised entities can apply their own resolution step on top of these arrays.

Table 11: Article-level descriptive statistics of the deposited CCF Database. Mean share is the average per-article proportion of sentences positive for the category; the top-frame column reports the percentage of articles for which the frame has the highest share among the eight frames. Computed from `CCF_full_data`, `CCF_processed_data`, and `CCF_article_aggregates`.

Corpus: 283,964 articles, 9,908,776 two-sentence units;
736 words/article on average (SD = 536; median = 680).

	Main frames		**Primary categories & tone**	
	Mean share (%)	**Top frame (%)**		**Mean share (%)**
Political frame	35.4	53.7	Messenger present	52.9
Economic frame	17.6	21.0	Event present	23.0
Scientific frame	7.9	9.4	Solution present	27.2
Environmental frame	3.6	4.5	Canadian context	46.7
Justice frame	3.6	1.4	Urgency to act	2.6
Cultural frame	3.7	7.2	Positive tone	9.6
Health frame	1.7	1.7	Negative tone	39.6
Security frame	0.5	0.4	Neutral tone	62.7

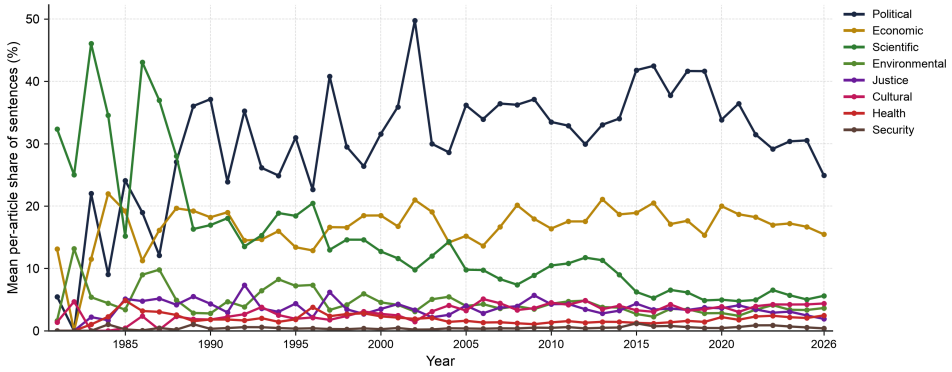

Figure 7: Mean per-article share of each main frame, year by year, 1980–2026. Each point is the mean of the corresponding `prop_X` column of `CCF_article_aggregates` across the articles published in that year, joined with `CCF_full_data` on `doc_id`. The pre-1988 portion of the curve is computed on fewer than fifty articles per year.

Table 12: Ten most frequently mentioned entities of each type in the deposited CCF Database. "Articles" is the number of articles in which the entity appears at least once. Computed from the JSONB arrays of `CCF_article_entities` (`entities_per`, `entities_org`, `entities_loc`).

	Persons (PER)		**Organisations (ORG)**		**Locations (LOC)**	
#	**Entity**	**Articles**	**Entity**	**Articles**	**Entity**	**Articles**
1	Justin Trudeau	24,996	NDP	21,366	Canada	136,896
2	Stephen Harper	24,174	United Nations	13,632	U.S.	59,003
3	Trudeau	20,909	UN	13,158	Ottawa	49,617
4	Harper	20,069	OTTAWA	9,940	Alberta	47,836
5	Donald Trump	13,486	Star	9,836	United States	38,626
6	Barack Obama	11,936	European Union	8,011	Ontario	36,709
7	Trump	10,417	Environment Canada	7,967	Toronto	35,531
8	Obama	8,660	House of Commons	7,519	Quebec	23,264
9	George W. Bush	7,282	National Post	6,969	U.S	22,501
10	Bush	6,885	Senate	6,953	Québec	21,761

Technical Validation

Technical validation of the CCF database and the trained models proceeded in three steps. First, a stratified validation set was constructed in line with established best practices for multi-label classification. Second, an expert annotator in climate change politics produced the reference annotations that define the gold standard used for all model evaluation. Third, an independent second annotator replicated the coding following a structured protocol comprising a training phase (sentences 1–600) and a blind coding phase (sentences 601–1000), enabling assessment of inter-coder reliability. The 1,000-sentence gold standard is strictly disjoint from the 4,000-sentence training/validation pool used in Phase 3. We explicitly excluded all sentence identifiers present in the training data before drawing the stratified sample, so that no test sentence was ever seen by a trained model at training time.

Construction of the stratified validation set

The validation set comprised 1,000 sentences (500 per language) extracted from the fully annotated database using a root-inverse probability weighting scheme combined with strict structural constraints, summarised in Figure 8. Each candidate sentence i received a sampling weight

$$w_i \;\propto\; \sum_{j \in C(i)} \frac{1}{\sqrt{f_j}}, \tag{1}$$

where $C(i)$ denotes the set of categories that are positive for sentence i and f_j denotes the corpus-wide prevalence of category j. The square root in the denominator deliberately moderated the influence of extremely rare categories, while still privileging them over common ones. Concretely, a sentence annotated with both *Economic Frame* (15.4% prevalence in the corpus) and *Loss of Indigenous Practices* (0.28% prevalence) received a weight of

$$\frac{1}{\sqrt{0.154}} + \frac{1}{\sqrt{0.0028}} \;=\; 2.55 + 18.90 \;=\; 21.45,$$

which was roughly eight times the weight of a sentence containing only *Economic Frame*, and ensured that under-represented categories nevertheless reached the validation set.

Two structural constraints were then applied on top of this weighted draw. First, a floor of 40 positive examples per category was targeted at sampling time—counted on the *model-predicted* labels, since sampling necessarily precedes the expert coding of the gold standard—so that each category would enter the validation set with as much statistical support as its corpus prevalence allowed. This target could not always be met for rare categories, and even where 40 predicted positives were drawn, the number of *expert-confirmed* positives can be substantially lower: the effective support ranges from 2 (*Disruption of military operations*, whose corpus prevalence is on the order of 10^{-5}) to several hundred, and 23 of the 65 categories fall below 40 confirmed positives. The per-category support is reported in Supplementary Table S7 so that readers can weigh each per-category estimate accordingly. Second, a 35% maximum allocation per category was imposed to prevent any single prevalent category from monopolising the validation set; without this ceiling, *Actors/Messengers Detection* alone, present in 48.8% of database sentences, could plausibly have dominated the entire sample and left insufficient space for the rarer categories.

Expert gold-standard annotation and model performance

All 1,000 sentences in the validation set were manually annotated by an expert annotator specialising in climate change politics. These expert annotations defined the gold standard against which all model performance metrics reported in this section were computed.

Model performance against the gold standard is reported in Table 13 and is summarised through two complementary quantities, both reproducible from the deposited per-category metrics (`Database/Training_data/final_annotation_metrics_normalized.csv`, Supplementary Table S7). The *pooled* macro F1 aggregates the predictions of all 65 categories over the 1,000 gold-standard sentences into a single confusion matrix before computing the score: it reaches 0.866 overall (0.869 for English, 0.864 for French) and measures the quality of the annotation layer as a whole, weighting every annotation decision equally. The *mean per-category* macro F1—the unweighted average of the 65 per-category macro F1 scores, i.e. the same aggregation as the training-phase figure of 0.826 reported in Phase 3—is 0.773. The like-for-like comparison is therefore 0.826 in training versus 0.773 on the gold standard: the ordinary out-of-sample decrease expected when models face an independent, disjoint sample, concentrated in the low-support categories whose estimates are noisiest. Two design features of the validation set should be kept in mind when reading the per-category values: the sampling deliberately over-represents sentences positive for low-prevalence categories (the root-inverse weighting above), which are precisely the categories targeted by the reinforced training protocol, so post-reinforcement gains are more visible here than in the training-time average; and the per-category support varies widely (see above), so estimates for low-support categories carry wide uncertainty. As detailed in Supplementary Table S7, primary detection categories achieved the highest performance, with *Canadian Context* (F1=0.982) and *Actors/Messengers* (F1=0.961) showing near-perfect classification. The frames maintained strong performance despite their semantic complexity, with *Scientific Frame* (F1=0.890) and *Environmental Frame* (F1=0.871) leading this category. The hierarchical classification strategy proved particularly effective, as evidenced by consistently higher recall scores for detection categories (mean recall=0.891) that triggered conditional sub-category evaluation. Notably, the rarest category in the validation set, *Disruption of military operations*, reached

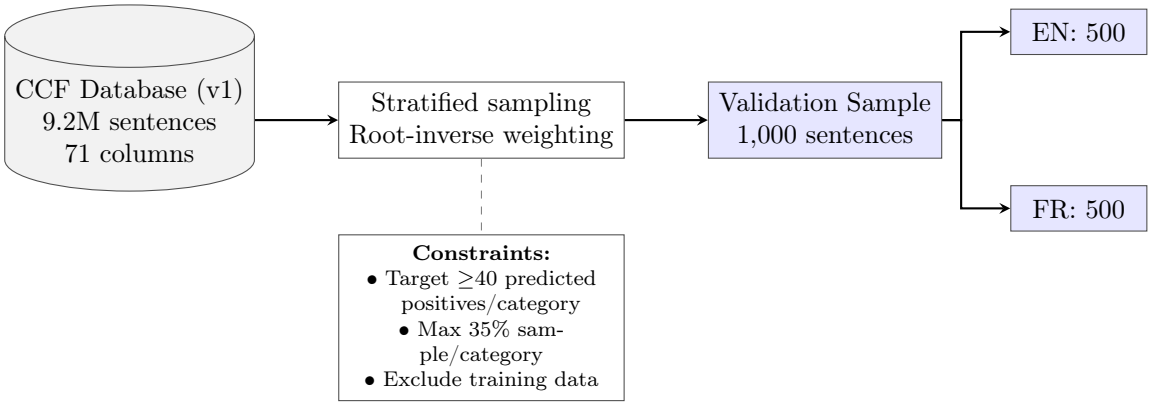

Figure 8: Stratified sampling procedure for final validation. Root-inverse probability weighting ensures balanced representation across all 65 annotation categories.

perfect precision on the French subset where two positive examples were available; the broader *Post-disaster military assistance* category, by contrast, had too few positives to support a reliable evaluation in either language and was flagged as exploratory in the reliability tiers (Supplementary Table S10).

To assess temporal stability, the gold-standard sentences were grouped by publication decade and the pooled macro F1 was recomputed within each period (Figure 9). Across the well-populated decades (1990s–2020s), pooled macro F1 ranges from 0.861 to 0.882—a maximum inter-decade spread of 2.1 points—with no monotonic degradation over time; the 1980s estimate (0.848) rests on only three gold-standard sentences and is shown for completeness. This temporal consistency supported the models' generalisation capacity across the full 48-year corpus span, and the per-decade computation is fully reproducible from the deposited gold standard and annotations (`Results/Scripts/2_temporal_f1_validation.py`).

Table 13: Overall validation performance metrics

Language	**F1 Macro**	**F1 Micro**	**F1 Weighted**
English (EN)	0.869	0.909	0.911
French (FR)	0.864	0.902	0.905
Combined (ALL)	**0.866**	**0.905**	**0.908**

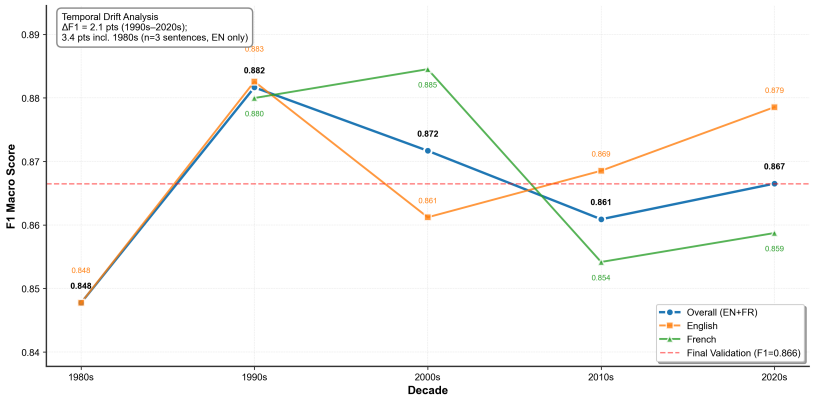

Figure 9: Temporal validation: pooled macro F1 of the deployed models against the gold standard, by publication decade of the validated sentences. Variation across the well-populated decades (1990s–2020s) stays within 2.1 points, with no monotonic degradation; the 1980s point rests on three sentences only.

Inter-coder assessment of the gold standard

Inter-coder reliability assessment was completed with a second independent annotator evaluating the same 1,000 randomly selected sentences from the validation set. The protocol followed a structured approach: during the first 600 sentences, the expert annotator provided training to the second coder, explaining category definitions and answering questions without directly intervening in the coding process. This training phase allowed the second

coder to progressively learn the annotation framework while maintaining independence in coding decisions. After sentence 600, the annotation process became completely blind, with no communication between coders. The resulting comparison between the second annotator's labels and the expert gold-standard annotations provided an explicit evaluation of the reliability and reproducibility of the gold standard.

Given that the authors' main interest lay in the level of reliability once the codebook had been learned, the metrics for the blind coding phase (sentences 601–1000) are reported as the primary benchmark. Table 14 summarises the agreement between the second annotator and the expert gold standard across all 65 annotation categories in this phase. These values indicated moderate agreement on Cohen's κ and substantial agreement on Gwet's AC1, while Krippendorff's α exceeded the commonly used 0.667 threshold for acceptable reliability. In other words, the gold-standard annotations produced by the expert could be reproduced to a reasonably high degree by an independent trained annotator, even in a setting with 65 multi-label categories and often dense and multi-topic climate discourse.

However, several limitations should be kept in mind. The moderate κ values reflected the inherent difficulty of climate discourse annotation at the sentence level, where many sentences contain subtle framings and overlapping categories. Chance-corrected indices such as Cohen's κ are also known to be sensitive to category imbalance—the well-documented *kappa paradox*, whereby high raw agreement can coexist with low κ when a category's prevalence is very low or very high [42, 43]. Prevalence imbalance is pronounced in this setting, and the substantially higher Gwet's AC1—an index specifically designed to resist this paradox—suggested that part of the apparent disagreement in κ was driven by prevalence rather than by systematic confusion between coders; the categories concerned are flagged `paradox_kappa`

Table 14: Inter-coder reliability metrics for the blind coding phase (sentences 601–1000)

	Cohen's κ	Gwet's AC1	Krippendorff's α
Blind coding phase (601–1000, $n = 400$)	0.596	0.894	0.698

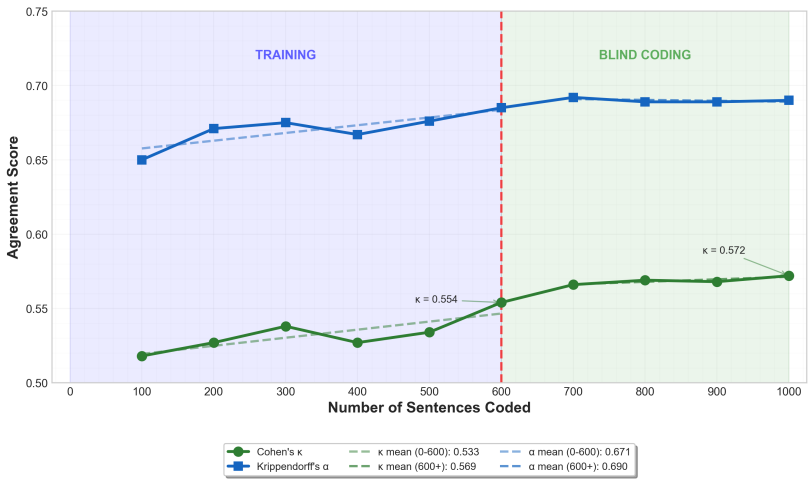

Figure 10: Inter-coder reliability progression across 1,000 annotated sentences, computed in successive blocks of 100 sentences.

in the reliability lookup. A further limitation concerns the training data itself: as discussed in Phase 3, all 4,000 training sentences were coded by a single expert, with the second coder reserved for the independent gold-standard validation. This design is consistent with the standard recommendation of stabilising the codebook through one expert before validating on a held-out sample [38], but it nevertheless means that the model behaviour reflects, at training time, the interpretive choices of a single coder. The per-category reliability tiers (Supplementary Table S10) and the blind-phase agreement metrics (Supplementary Table S9) are designed to make these choices auditable; future releases of the CCF will extend the multi-coder protocol to a stratified subsample of the training data. A second limitation concerns the three pre-trained NER models (Supplementary Table S6), whose precision, recall, and F1 figures are reported from the original-model documentation rather than from a CCF-specific gold standard. A systematic in-domain NER evaluation against a manually coded sample is planned for a subsequent release of the database; users are accordingly advised to treat the named-entity rollup as a high-recall extraction layer rather than as a benchmarked classifier.

To better understand how reliability evolved as the second annotator internalised the codebook, agreement metrics in successive blocks of 100 sentences were computed. This analysis showed a pattern consistent with a learning curve for a complex codebook: both Cohen's κ and Krippendorff's α increased over time, with higher values in the later blocks than in the early ones. Figure 10 illustrates this progression across the 1,000 sentences, with a continuing improvement after the transition to blind coding at sentence 600. The upward trajectory of both indices across the 100-sentence blocks indicated that, once training was complete, the annotation framework could be applied consistently. This reinforces the view that the expert-defined gold standard is reproducible by trained coders.

Validation of the database enrichment

Three integrity checks were run at the end of the enrichment pipeline to confirm that the derived tables remained exactly synchronised with the underlying sentence-level annotations.

First, the article-level aggregates were cross-validated against the source sentence table: for each of the 65 `prop_X` columns, the SQL identity `prop_X = SUM(COALESCE(X, 0)) / COUNT(*)` was verified to hold within floating-point tolerance (10^{-9}) for every one of the 283,964 articles, and the `top_frame` column was confirmed to equal the column code with the highest `prop_X` in every non-empty row. Second, the per-article entity rollup was checked by re-extracting the union of all distinct entities from the `ner_entities` JSON column for a stratified random sample of 1,000 articles and comparing the result element-wise against the JSONB arrays stored in `CCF_article_entities`; all 1,000 articles matched exactly. Third, the BGE-M3 embeddings were verified to have an L2 norm of 1.0 ± 10^{-3} on a random sample of 10,000 rows—the expected behaviour of vectors produced with `normalize_embeddings = True`—and an HNSW round-trip query (retrieve the top-1 nearest neighbour of an arbitrary vector and confirm that it is the vector itself, or an exact textual duplicate) returned the probe at cosine distance below 10^{-3} in 99 of 100 spot-checks with the search parameter `hnsw.ef_search` raised to 400—the expected behaviour of an approximate index, whose recall is governed by the query-time search width.

Usage Notes

Start from the reliability tiers. The recommended workflow for any substantive analysis is to open `CCF_reliability_tiers` first and restrict the analysis to Tier A categories (27 of 65), unless lower-tier categories are explicitly treated as exploratory. Because categories are *columns* in the annotation tables, the lookup is used to select columns rather than to join rows. In SQL, against the restored PostgreSQL edition:

```
- Tier-A category codes
SELECT code FROM "CCF_reliability_tiers" WHERE tier_overall = 'A';

- Example:  average attention to the political frame (a Tier-A category)
- across articles, by year
SELECT EXTRACT(YEAR FROM f.date) AS year, AVG(a.prop_political_frame)
FROM "CCF_article_aggregates" a
JOIN "CCF_full_data" f USING (doc_id)
GROUP BY year ORDER BY year;
```

The same selection in Python, against the Parquet mirror:

```
import pandas as pd
tiers = pd.read_parquet("CCF_reliability_tiers.parquet")
keep = ["prop_" + c for c in tiers.loc[tiers.tier_overall.eq("A"),
"code"]]
agg = pd.read_parquet("CCF_article_aggregates.parquet",
    columns=["doc_id", *keep])
```

Language coverage of two categories. *Post-disaster military assistance* and *disruption of military operations* are annotated on French sentences only (no English model could be trained; see Phase 3): their values are `NULL` on English rows of `CCF_processed_data` and coalesced to zero in the aggregates. Analyses of these two categories should be restricted to the French subset.

Versioning and updates. The database is refreshed approximately every six months through the CCF Observatory's continuous extraction and annotation pipeline. For reproducible work, cite and pin the *version* DOI of the release used (this paper describes v2.1.0); use the concept DOIs to always obtain the latest release. Release notes in each deposit's README document the corpus growth between versions; the validation layer (training data, gold standard, reliability tiers) is frozen and carries over unchanged between releases unless explicitly re-run and re-documented.

Data Availability

The CCF Database v2.1.0 is openly available on Zenodo under a CC-BY 4.0 licence, without embargo or access restriction, in two editions sharing identical schemas: the canonical PostgreSQL edition (DOI 10.5281/zenodo.22019287 [44]) and the Apache Parquet mirror (DOI 10.5281/zenodo.22019288 [45]). The concept DOIs 10.5281/zenodo.20346363 and 10.5281/zenodo.20346372 always resolve to the latest release.

Code Availability

All custom code used to build, validate, enrich, and export the database—the annotation pipeline (`Scripts/Annotation/`, scripts 01–17), the enrichment and export pipeline (`Scripts/Database_creation/`), and the manuscript's table- and figure-generation scripts—is openly archived on Zenodo (DOI 10.5281/zenodo.22019289 [46]; concept DOI 10.5281/zenodo.21921214 always resolves to the latest version) and, for convenience, bundled with both data editions as `ccf_code_and_paper.tar.gz`. A development mirror is maintained at https://github.com/antoinelemor/CCF-canadian-climate-framing. Pre-existing third-party models and libraries used by the pipeline (BERT-base-cased, CamemBERT-base, the AugmentedSocialScientist fork, `dslim/bert-base-NER`, `Jean-Baptiste/camembert-ner`, spaCy `fr_core_news_lg`, BAAI `bge-m3`, `pgvector`, DuckDB) are identified, with citations, in the Methods section.

Authors Contribution Statement

Antoine Lemor designed and implemented the complete data architecture, including data extraction, cleaning, parsing, and database construction. Developed all machine learning models and methodological approaches. Wrote all code and conducted quantitative analyses. Drafted the methodology and results sections.

Alizée Pillod developed the theoretical and conceptual foundations underlying the annotation framework. Designed the complete annotation scheme and performed all manual annotation. Drafted the theoretical background and literature review sections.

Matthew Taylor contributed to conceptual development and analysis design. Verified statistical analyses. Reviewed and edited the complete manuscript.

Richard Nadeau contributed to the conceptual development of the project.

Acknowledgements

The authors thank Maëlle Rouyer for serving as the second annotator on the inter-coder reliability evaluation, and Professors Vincent Arel-Bundock, Evelyne Brie, and Jean-François Godbout for their generous comments and the time they devoted to supporting this research.

Funding Information

Funding for this research was provided by the Fonds de recherche du Québec (FRQ grant number 337081) and the Canadian Social Sciences and Humanities Research Council (SSHRC grant number RNH01687).

References

[1] Dietram A. Scheufele and David Tewksbury. "Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models". en. In: *Journal of Communication* 57.1 (Mar. 2007). Publisher: John Wiley & Sons, Ltd, pp. 9–20. ISSN: 1460-2466. DOI: 10.1111/j.0021-9916.2007.00326.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.0021-9916.2007.00326.x.

[2] Stefaan Walgrave and Peter Van Aelst. "The Contingency of the Mass Media's Political Agenda Setting Power: Toward a Preliminary Theory". en. In: *Journal of Communication* 56.1 (Mar. 2006), pp. 88–109. ISSN: 0021-9916, 1460-2466. DOI: 10.1111/j.1460-2466.2006.00005.x. URL: https://academic.oup.com/joc/article/56/1/88-109/4102573.

[3] Stanley Cohen. *Folk Devils and Moral Panics: The Creation of the Mods and Rockers*. London: MacGibbon & Kee, 1972.

[4] William Hoynes. "Media and the Construction of Social Problems". In: *The Cambridge Handbook of Social Problems*. Ed. by A. Javier Treviño. Vol. 1. Cambridge: Cambridge University Press, 2018, pp. 477–496. ISBN: 978-1-108-42616-9. DOI: 10.1017/9781108656184.027. URL: https://www.cambridge.org/core/books/cambridge-handbook-of-social-problems/media-and-the-construction-of-social-problems/70A9AFB47F08651CAB1B944034BB18FA.

[5] Marco Toledo Bastos. "Shares, Pins, and Tweets: News readership from daily papers to social media". In: *Journalism Studies* 16.3 (May 2015). Publisher: Routledge _eprint: https://doi.org/10.1080/1461670X.2014.891857, pp. 305–325. ISSN: 1461-670X. DOI: 10.1080/1461670X.2014.891857. URL: https://doi.org/10.1080/1461670X.2014.891857.

[6] Andreu Casero-Ripolles. *Impact of COVID-19 on the Media System. Communicative and Democratic Consequences of News Consumption during the Outbreak*. en. SSRN Scholarly Paper. Rochester, NY, Apr. 2020. URL: https://papers.ssrn.com/abstract=3594133.

[7] N. Lulham et al. *Canada in a changing climate: synthesis report*. English. Tech. rep. Natural Resources Canada, Dec. 2023. URL: https://ostrnrcan-dostrncan.canada.ca/handle/1845/271933.

[8] Natural Resources Canada. *Energy Fact Book, 2025-2026*. Tech. rep. Natural Resources Canada, 2025. URL: https://energy-information.canada.ca/sites/default/files/2025-11/energy-fact-book-2025-2026.pdf.

[9] Kathryn Harrison. "The Struggle of Ideas and Self-Interest in Canadian Climate Policy". en. In: *Global Commons, Domestic Decisions: The Comparative Politics of Climate Change*. Ed. by Kathryn Harrison and Lisa McIntosh Sundstrom. Google-Books-ID: VsTxCwAAQBAJ. MIT Press, July 2010, p. 336. ISBN: 978-0-262-28887-3.

[10] C. Hatch et al. *What Do Canadians Really Think About Climate Change? A Summary of Public Opinion Research for Communicators*. Tech. rep. Re.Climate, 2025. URL: https://reclimate.ca/wp-content/uploads/2025/06/Re.Climate-Public-Opinion-Summary-2025-Report.pdf.

[11] Nathan Young and Eric Dugas. "Comparing Climate Change Coverage in Canadian English- and French-Language Print Media: Environmental Values, Media Cultures, and the Narration of Global Warming". In: *The Canadian Journal of Sociology / Cahiers canadiens de sociologie* 37.1 (2012). Publisher: Canadian Journal of Sociology, pp. 25–54. ISSN: 0318-6431. DOI: 10.29173/cjs9733. URL: https://doi.org/10.29173/cjs9733.

[12] Nia King et al. "How do Canadian media report climate change impacts on health? A newspaper review". en. In: *Climatic Change* 152.3 (Mar. 2019), pp. 581–596. ISSN: 1573-1480. DOI: 10.1007/s10584-018-2311-2. URL: https://doi.org/10.1007/s10584-018-2311-2.

[13] Pradip Swarnakar and Ashutosh Modi. *NLP for Climate Policy: Creating a Knowledge Platform for Holistic and Effective Climate Action*. arXiv:2105.05621 [cs]. May 2021. DOI: 10.48550/arXiv.2105.05621. URL: http://arxiv.org/abs/2105.05621.

[14] Mike S. Schäfer and Valerie Hase. "Computational methods for the analysis of climate change communication: Towards an integrative and reflexive approach". en. In: *Wiley Interdisciplinary Reviews: Climate Change* 14.2 (Mar. 2023). Publisher: John Wiley & Sons, Ltd, e806. ISSN: 1757-7799. DOI: 10.1002/wcc.806. URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcc.806.

[15] Simon David Hirsbrunner. "Computational methods for climate change frame analysis: Techniques, critiques, and cautious ways forward". en. In: *Wiley Interdisciplinary Reviews: Climate Change* 15.5 (Sept. 2024). Publisher: John Wiley & Sons, Ltd, e902. ISSN: 1757-7799. DOI: 10.1002/wcc.902. URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcc.902.

[16] Tom Nicholls and Pepper D. Culpepper. "Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities". In: *Computational Political Communication*. Ed. by Yannis Theocharis and Andreas Jungherr. Num Pages: 23. Routledge, 2024.

[17] Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. "Media Framing: A typology and Survey of Computational Approaches Across Disciplines". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15407–15428. DOI: 10.18653/v1/2024.acl-long.822. URL: https://aclanthology.org/2024.acl-long.822/.

[18] Robert M. Entman. "Framing: Toward Clarification of a Fractured Paradigm". en. In: *Journal of Communication* 43.4 (Dec. 1993), pp. 51–58. ISSN: 0021-9916, 1460-2466. DOI: 10.1111/j.1460-2466.1993.tb01304.x. URL: https://academic.oup.com/joc/article/43/4/51-58/4160153.

[19] Yiwei Luo, Dallas Card, and Dan Jurafsky. "Detecting Stance in Media on Global Warming". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 3296–3315. DOI: 10.18653/v1/2020.findings-emnlp.296. URL: https://aclanthology.org/2020.findings-emnlp.296/.

[20] Paul Meddeb et al. "Counteracting French Fake News on Climate Change Using Language Models". en. In: *Sustainability* 14.18 (Sept. 2022). Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher, p. 11724. ISSN: 2071-1050. DOI: 10.3390/su141811724. URL: https://www.mdpi.com/2071-1050/14/18/11724.

[21] Lea Frermann et al. "Conflicts, Villains, Resolutions: Towards models of Narrative Media Framing". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 8712–8732. DOI: 10.18653/v1/2023.acl-long.486. URL: https://aclanthology.org/2023.acl-long.486/.

[22] Nicholas Badullovich et al. "Manual and automatic paragraph-level analysis of climate change framing in academic journal editorials". In: *Scientometrics* 131.1 (2026), pp. 353–387. ISSN: 0138-9130. DOI: 10.1007/s11192-025-05494-w. URL: https://doi.org/10.1007/s11192-025-05494-w.

[23] Newspapers Canada. *Circulation Report: Daily Newspapers*. Tech. rep. Newspapers Canada, 2015. URL: https://nmc-mic.ca/wp-content/uploads/2016/06/2015-Daily-Newspaper-Circulation-Report-REPORT_FINAL.pdf.

[24] Regina G. Lawrence. "Framing Obesity: The Evolution of News Discourse on a Public Health Issue". EN. In: *Harvard International Journal of Press/Politics* 9.3 (July 2004). Publisher: SAGE Publications, pp. 56–75. ISSN: 1081-180X. DOI: 10.1177/1081180X04266581. URL: https://doi.org/10.1177/1081180X04266581.

[25] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431. DOI: 10.18653/v1/E17-2068. URL: https://aclanthology.org/E17-2068/.

[26] Armand Joulin et al. "FastText.zip: Compressing text classification models". In: *arXiv preprint arXiv:1612.03651* (2016). URL: https://arxiv.org/abs/1612.03651.

[27] Reiner Grundmann. "Using large text news archives for the analysis of climate change discourse: some methodological observations". In: *Journal of Risk Research* 25.3 (Mar. 2022). Publisher: Routledge _eprint: https://doi.org/10.1080/13669877.2021.1894471, pp. 395–406. ISSN: 1366-9877. DOI: 10.1080/13669877.2021.1894471. URL: https://doi.org/10.1080/13669877.2021.1894471.

[28] Reiner Grundmann. "Climate change as a wicked social problem". en. In: *Nature Geoscience* 9.8 (Aug. 2016). Publisher: Nature Publishing Group, pp. 562–563. ISSN: 1752-0908. DOI: 10.1038/ngeo2780. URL: https://www.nature.com/articles/ngeo2780.

[29] Manfred Stede and Ronny Patz. "The Climate Change Debate and Natural Language Processing". In: *Proceedings of the 1st Workshop on NLP for Positive Impact*. Ed. by Anjalie Field et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 8–18. DOI: 10.18653/v1/2021.nlp4posimpact-1.2. URL: https://aclanthology.org/2021.nlp4posimpact-1.2/.

[30] James Ford and Diana King. "Coverage and framing of climate change adaptation in the media: A review of influential North American newspapers during 1993–2013". In: *Environmental Science & Policy* 48 (Apr. 2015), pp. 137–146. ISSN: 1462-9011. DOI: 10.1016/j.envsci.2014.12.003. URL: https://www.sciencedirect.com/science/article/pii/S1462901114002317.

[31] Craig Trumbo. "Constructing climate change: claims and frames in US news coverage of an environmental issue". en. In: *Public Understanding of Science* 5.3 (July 1996), pp. 269–283. ISSN: 0963-6625. DOI: 10.1088/0963-6625/5/3/006. URL: https://doi.org/10.1088/0963-6625/5/3/006.

[32] Maxwell T. Boykoff. *Who Speaks for the Climate?: Making Sense of Media Reporting on Climate Change*. en. Google-Books-ID: zsdRZR88tdkC. Cambridge University Press, Sept. 2011. ISBN: 978-1-139-50179-8.

[33] Anthony Downs. "Up and down with ecology—the issue attention cycle". In: *Public Interest* 28 (1972), pp. 38–50. URL: https://fbaum.unc.edu/teaching/articles/Downs_Public_Interest_1972.pdf.

[34] Xinsheng Liu, Eric Lindquist, and Arnold Vedlitz. "Explaining Media and Congressional Attention to Global Climate Change, 1969-2005: An Empirical Test of Agenda-Setting Theory". en. In: *Political Research Quarterly* 64.2 (2011), pp. 405–419. DOI: 10.1177/1065912909346744. URL: https://journals.sagepub.com/doi/epdf/10.1177/1065912909346744.

[35] Susanne C. Moser. "Reflections on climate change communication research and practice in the second decade of the 21st century: what more is there to say?" en. In: *Wiley Interdisciplinary Reviews: Climate Change* 7.3 (May 2016). Publisher: John Wiley & Sons, Ltd, pp. 345–369. ISSN: 1757-7799. DOI: 10.1002/wcc.403. URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcc.403.

[36] Sefat Salama and Khalil Aboukoura. "Role of Emotions in Climate Change Communication". en. In: *Handbook of Climate Change Communication: Vol. 1: Theory of Climate Change Communication*. Ed. by Walter Leal Filho et al. Cham: Springer International Publishing, 2018, pp. 137–150. ISBN: 978-3-319-69838-0. DOI: 10.1007/978-3-319-69838-0_9. URL: https://doi.org/10.1007/978-3-319-69838-0_9.

[37] Robin L. Nabi, Abel Gustafson, and Risa Jensen. "Framing Climate Change: Exploring the Role of Emotion in Generating Advocacy Behavior". EN. In: *Science Communication* 40.4 (Aug. 2018). Publisher: SAGE Publications Inc, pp. 442–468. ISSN: 1075-5470. DOI: 10.1177/1075547018776019. URL: https://doi.org/10.1177/1075547018776019.

[38] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. 4th. Thousand Oaks, CA: SAGE Publications, 2019. ISBN: 978-1-5063-9566-1.

[39] Salomé Do, Étienne Ollion, and Rubing Shen. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy". en. In: *Sociological Methods & Research* 53.3 (2024). Publisher: SAGE Publications Inc. First published online 9 December 2022., pp. 1167–1200. ISSN: 0049-1241. DOI: 10.1177/00491241221134526. URL: https://doi.org/10.1177/00491241221134526.

[40] Antoine Lemor. *antoinelemor/AugmentedSocialScientistFork*. original-date: 2025-04-26T21:39:03Z. June 2025. URL: https://github.com/antoinelemor/AugmentedSocialScientistFork.

[41] Jianlv Chen et al. "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2318–2335. DOI: 10.18653/v1/2024.findings-acl.137. URL: https://aclanthology.org/2024.findings-acl.137/.

[42] Alvan R. Feinstein and Domenic V. Cicchetti. "High agreement but low kappa: I. The problems of two paradoxes". In: *Journal of Clinical Epidemiology* 43.6 (1990), pp. 543–549. DOI: 10.1016/0895-4356(90)90158-L. URL: https://doi.org/10.1016/0895-4356(90)90158-L.

[43] Ted Byrt, Janet Bishop, and John B. Carlin. "Bias, prevalence and kappa". In: *Journal of Clinical Epidemiology* 46.5 (1993), pp. 423–429. DOI: 10.1016/0895-4356(93)90018-V. URL: https://doi.org/10.1016/0895-4356(93)90018-V.

[44] Antoine Lemor et al. *CCF Database (PostgreSQL edition): A sentence-level corpus of 283,964 Canadian climate-change articles from 22 news outlets (1978–2026)*. Version 2.1.0. Canonical PostgreSQL `pg_dump` archive with `pgvector` extension and HNSW cosine indexing. Zenodo, 2026. DOI: 10.5281/zenodo.22019287. URL: https://doi.org/10.5281/zenodo.22019287.

[45] Antoine Lemor et al. *CCF Database (Apache Parquet mirror): column-oriented edition of the Canadian Climate Framing corpus (1978–2026)*. Version 2.1.0. Column-oriented Apache Parquet mirror of the canonical PostgreSQL edition; one ZSTD-compressed file per table, readable by pandas, polars, R/arrow, DuckDB, and Spark without a database server. Zenodo, 2026. DOI: 10.5281/zenodo.22019288. URL: https://doi.org/10.5281/zenodo.22019288.

[46] Antoine Lemor et al. *CCF – Canadian Climate Framing: Code, documentation and reproducible pipeline*. Version 2.1.0. Complete annotation, enrichment, and export pipeline of the CCF Database, with documentation and the expert-coded training and gold-standard sentences. 2026. DOI: 10.5281/zenodo.22019289. URL: https://doi.org/10.5281/zenodo.22019289.

CCF Database: A Machine-Learning-Annotated Corpus of 283,964 Canadian Climate Articles (1978–2026)

Supplementary Information

Antoine Lemor*[1], Alizée Pillod†[2], Matthew Taylor‡[2], and Richard Nadeau§[2]

[1]Université de Sherbrooke
[2]Université de Montréal

This document collects the Supplementary Information accompanying the CCF Database manuscript. It contains: prior-literature and corpus composition tables (Supplementary Tables S1 and S2), the complete annotation framework with operational definitions (Supplementary Table S3), the complete per-category training metrics and dataset distribution (Supplementary Tables S4–S5), the Named Entity Recognition performance summary (Supplementary Table S6), the per-category validation metrics on the 1,000-sentence gold standard (Supplementary Table S7), the per-category distribution of annotations across the database (Supplementary Table S8), and four additional tables added during revision: per-category inter-coder reliability (Supplementary Table S9), reliability tier assignment (Supplementary Table S10), per-model training provenance (Supplementary Table S11), and the complete data dictionary of the enriched database (Supplementary Table S12).

Reproducibility note. Tables S1, S2, S3, and S6 are static prose / lookup tables inlined directly in this LaTeX source (their content does not depend on the live data). Tables S4 to S12 are reproducibly generated by `Scripts/Annotation/17_generate_tables.py` from the canonical CSVs in the `Database/Training_data/` directory and the live `CCF_Database` (Table S8 only); the generator reuses the category-normalisation API exposed by `Scripts/Annotation/15_normalize_categories.py`. Re-running scripts 15, 16, and 17 followed by a fresh `pdflatex` pass regenerates every numerical value in this document from the underlying data.

*antoine.lemor@umontreal.ca
†alizee.pillod@umontreal.ca
‡matthew.taylor@umontreal.ca
§richard.nadeau@umontreal.ca

Supplementary Table S1. Previous studies on climate change coverage in Canadian newspapers.

Study	Newspaper(s)	Scope	Language	Time Period
Good [1]	The Globe and Mail	National	English	2007
	The National Post	National	English	
	The Toronto Star	National	English	
	The Calgary Herald	Regional	English	
	The Edmonton Journal	Regional	English	
	The Chronicle Herald	Regional	English	
	The Montreal Gazette	Regional	English	
	The Charlottetown Guardian	Regional	English	
	Halifax Daily News	Regional	English	
	The Ottawa Citizen	Regional	English	
	The Star Phoenix	Regional	English	
	The Telegram	Regional	English	
	The Times Colonist	Regional	English	
	The Vancouver Sun	Regional	English	
	The Winnipeg Sun	Regional	English	
	Yukon News	Regional	English	
Rowe [2]	The Globe and Mail	National	English	2007–2008
	The Toronto Star	National	English	
	The Calgary Herald	Regional	English	
	The Toronto Sun	Regional	English	
	The Montreal Gazette	Regional	English	
	The Ottawa Citizen	Regional	English	
Young and Dugas [3]	The Globe and Mail	National	English	1988–2008
	The National Post	National	English	
Anne DiFrancesco and Young [4]	The Globe and Mail	National	English	2008
	The National Post	National	English	
Ahchong and Dodds [5]	The Globe and Mail	National	English	1988–2007
	The Toronto Star	National	English	
Young and Dugas [6]	The Globe and Mail	National	English	2007–2008
	The National Post	National	English	
	The Toronto Star	National	English	

Continued on next page

Supplementary Table S1 – *Continued from previous page*

Study	Newspaper(s)	Scope	Language	Time Period
	The Vancouver Sun	Regional	English	
	The Calgary Herald	Regional	English	
	The Telegram	Regional	English	
	Le Devoir	Regional	French	
	La Presse	Regional	French	
Stoddart and Tindall [7]	The Globe and Mail	National	English	1999–2010
	The National Post	National	English	
Ford and King [8]	The Globe and Mail	National	English	1993–2013
	The Toronto Star	National	English	
Stoddart and Smith [9]	The Globe and Mail	National	English	2006–2010
	The National Post	National	English	
Stoddart et al. [10]	The Globe and Mail	National	English	1997–2010
	The National Post	National	English	
Barkemeyer et al. [11]	The Globe and Mail	National	English	2008
	The National Post	National	English	
	The Toronto Star	National	English	
	The Vancouver Sun	Regional	English	
	The Toronto Sun	Regional	English	
King et al. [12]	The Globe and Mail	National	English	2005–2015
	The National Post	National	English	
	The Toronto Star	National	English	
	The Calgary Herald	Regional	English	
	The Chronicle Herald	Regional	English	
	The Vancouver Sun	Regional	English	
	The Whitehorse Daily Star	Regional	English	
	Journal de Montréal	Regional	French	
Pillod [13]	The Globe and Mail	National	English	2008–2020
Brandt [14]	The Toronto Star	National	English	2015–2018
	The Calgary Herald	Regional	English	
	The Vancouver Sun	Regional	English	
	The Chronicle Herald	Regional	English	

Continued on next page

Supplementary Table S1 – *Continued from previous page*

Study	**Newspaper(s)**	**Scope**	**Language**	**Time Period**
	The Winnipeg Free Press	Regional	English	
	Times and Transcript	Regional	English	
Hase et al. [15]	The Globe and Mail	National	English	2006–2018
	The Toronto Star	National	English	
Sachdeva and McCaffrey [16]	n/a	n/a	English	1986–2016
Stoddart and Yang [17]	The Globe and Mail	National	English	2015–2020
	The National Post	National	English	
	The Telegraph	Regional	English	
	The Telegram	Regional	English	
	The Chronicle Herald	Regional	English	
	The Guardian	Regional	English	
Chen et al. [18]	n/a	n/a	English	2018–2021

Note: This table summarizes the newspaper sources, geographic scope, language, and time periods covered by previous studies on climate change media coverage in Canada. Most studies focus exclusively on English-language national newspapers, particularly The Globe and Mail and The National Post.

Supplementary Table S2. Newspapers included in the CCF database.

Province/Territory	Newspaper	Language	Avg. Weekday Circ. (2015)	Articles
National	The Globe and Mail	English	226,171	31,127
	The National Post	English	79,157	20,829
	The Toronto Star	English	192,084	48,616
	CBC.ca[b]	English	–	2,941
	Radio-Canada.ca[b]	French	–	3,054
Alberta	The Edmonton Journal	English	59,244	18,763
	The Calgary Herald	English	58,865	20,349
British Columbia	The Vancouver Sun	English	78,901	19,068
	The Times Colonist	English	47,556	12,111
Manitoba	The Winnipeg Free Press	English	67,503	13,129
New Brunswick	L'Acadie Nouvelle[c]	French	18,102	5,143
Newfoundland & Labrador	The Telegram	English	15,380	6,208
Nova Scotia	The Chronicle Herald	English	71,304	11,018
Ontario	The Toronto Sun	English	85,775	3,333
	Le Droit[c]	French	28,616	4,728
Quebec	Le Devoir[c]	French	28,729	13,685
	Le Journal de Montréal	French	175,220	5,790
	La Presse	French	98,657	6,917
	La Presse+[a]	French	–	11,636
	The Montreal Gazette	English	53,344	9,799
Saskatchewan	The Star Phoenix	English	31,081	8,018
Yukon	The Whitehorse Daily Star[c]	English	945	7,702
Total				**283,964**

[a] In 2013, La Presse launched a free digital tablet edition called La Presse+ which offers slightly different content. The corpus considers articles published both on the website and in the digital tablet edition. Duplicates were discarded.
[b] News websites of the Canadian public broadcaster, added to the corpus by the CCF Observatory; their coverage begins in 2023. No print circulation applies.

[c] Coverage window bounded by the access period of the source aggregators: Le Droit until December 2023, The Whitehorse Daily Star until May 2024, Le Devoir and L'Acadie Nouvelle until March 2025. All other outlets are covered through July 2026 (corpus cut-off of release v2.1.0).

Supplementary Table S3. Complete CCF annotation framework.

#	Category	Code	Description
MAIN FRAMES			
Economic Frame			
1	**Economic Frame (Primary Category)**	`economic_frame`	If the sentence refers to climate change as an economic issue, including its negative or positive impacts on the economy, the financial costs or benefits of climate action, or the carbon footprint of the economic or industrial sector. This label captures any framing of climate change in economic terms, such as effects on jobs, markets, industries, or public and private finances.
2	Negative impacts of climate change on the economy	`eco_neg_impact`	If the sentence refers to the negative effects of climate change on the economy, such as damage to infrastructure, loss of agricultural productivity, disruptions to trade, increased costs for businesses, or economic decline in sectors vulnerable to climate shifts. This includes any economic losses directly resulting from changing climate conditions.
3	Positive impacts of climate change on the economy	`eco_pos_impact`	If the sentence refers to the positive effects of climate change on the economy, such as new agricultural opportunities, the opening of new trade routes, or the potential for economic growth in sectors adapting to climate shifts. This includes any economic gains directly resulting from changing climate conditions.
4	Economic disadvantages of climate action	`eco_cost`	If the sentence refers to the negative economic impacts caused by climate policies or climate action, such as costs associated with transitioning to green energy, investments in climate adaptation, debt incurred from financing climate action, or short-term job losses in fossil fuel industries. This category captures financial burdens and challenges linked to climate action or policy.
5	Economic benefits of climate action	`eco_benefit`	If the sentence refers to the positive economic impacts resulting from climate policies or climate action, such as job creation in renewable energy sectors, investments in green technologies, economic growth driven by sustainability initiatives, or cost savings from energy efficiency. This category captures financial gains and opportunities linked to climate action or policy.
6	Carbon footprint of the economic sector	`eco_footprint`	If the sentence refers to the environmental or carbon footprint of the economic or industrial sectors, including emissions and resource use related to manufacturing, transportation, energy production, mining, agriculture, and other commercial or industrial activities. This label covers the ecological impact generated by economic and industrial processes, not the effects of climate change on these sectors.
Health Frame			

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
7	**Health Frame (Primary Category)**	`health_frame`	If the sentence refers to climate change as a public health issue, including its negative or positive impacts on physical or mental health, the health co-benefits of climate action (e.g., cleaner air, active transportation), or the carbon footprint of the health sector. This label captures any framing of climate change in relation to health outcomes, healthcare systems, or health-related consequences and solutions.
8	Negative impacts of climate change on health	`health_neg_impact`	If the sentence refers to the negative effects of climate change on human health, such as increased risks of infectious diseases (e.g., malaria, dengue), respiratory and cardiovascular illnesses due to air pollution and heat exposure, heat-related illnesses (e.g., heatstroke), and mental health disorders (e.g., anxiety, depression) linked to environmental changes.
9	*Positive impacts of climate change on health**	`health_pos_impact`	If the sentence refers to the potential positive effects of climate change on human health, such as reductions in cold-related illnesses and deaths, or increased opportunities for outdoor activity due to milder winters.
10	Health co-benefits of climate action	`health_cobenefit`	If the sentence refers to health benefits resulting from climate policies or actions, such as improved air quality, better nutrition, reduced respiratory or cardiovascular diseases, fewer premature deaths, and other positive health outcomes linked to climate mitigation and adaptation efforts.
11	*Carbon footprint of the health sector**	`health_footprint`	If the sentence refers to the environmental or carbon footprint of the health sector, including emissions and resource use associated with healthcare facilities, medical equipment, pharmaceuticals production, patient and staff travel, and waste generated by medical activities. This label covers the ecological impact of delivering health services, not the impacts of climate change on health outcomes.
Security Frame			
12	**Security Frame (Primary Category)**	`security_frame`	If the sentence refers to the effects of climate change on security, such as violent conflict driven by resource scarcity, an influx of climate refugees, disruption of military operations, or the carbon footprint of the defense sector. This category includes broader security challenges arising from climate change, including geopolitical tensions or the strategic responses to climate impacts.
13	Presence of climate refugees	`security_refugees`	If the sentence refers to the displacement of populations or the presence of climate refugees as a result of a natural disaster, environmental degradation, or climate change impact. This category captures the movement of people fleeing climate-related disasters.

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
14	Conflict	`security_conflict`	If the sentence refers to the emergence or intensification of conflicts over natural resource exploitation due to climate change, such as disputes over water, land, or energy resources exacerbated by shifting climate patterns. This includes conflicts arising from scarcity or competition for increasingly stressed resources.
15	*Post-disaster military assistance**	`security_military`	If the sentence refers to the deployment of military forces as reinforcement after a natural disaster, such as wildfires, floods, or other extreme events. This includes military support in disaster response and recovery, like providing humanitarian aid or maintaining order during a crisis.
16	*Disruption of military operations**	`security_disruption`	If the sentence refers to the disruption of military operations due to a natural disaster affecting a military base or facilities. This includes cases where climate-induced events damage or disrupt military infrastructure, affecting defense readiness or response capabilities.
Justice Frame			
17	**Justice Frame (Primary Category)**	`justice_frame`	If the sentence frames climate change as a justice or moral issue, including references to who benefits or loses from climate action, differentiated responsibilities across countries or groups, unequal impacts of climate change, disparities in access to mitigation or adaptation measures, or concerns about intergenerational justice. This label captures ethical, equity-based, or fairness-oriented dimensions of climate discourse.
18	Winners and losers of climate action	`justice_winners`	If the sentence refers to how climate policies or actions create winners and losers, by benefiting certain groups (e.g., vulnerable populations, Indigenous communities, developing countries) while disadvantaging others (e.g., fossil fuel workers, carbon-intensive industries, or developed nations facing steeper transitions). This label focuses on distributional outcomes of climate policy.
19	Differentiated responsibility	`justice_responsibility`	If the sentence refers to the unequal responsibility for causing climate change, especially in reference to the principle of common but differentiated responsibilities. This includes statements emphasizing that developed countries, multinational corporations, or high-emitting individuals have historically contributed more to global warming and should therefore bear a greater share of the mitigation burden.
20	Unequal vulnerability to climate change	`justice_vulnerability`	If the sentence highlights how different populations are unequally affected by the impacts of climate change, due to geographic, social, or economic vulnerabilities. This includes references to marginalized groups such as women, seniors, racialized populations, people in developing countries, or those living in high-risk zones (e.g., coastal or arid regions). This label focuses on impacts, not responsibility or capacity.

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
21	Unequal access to climate action	`justice_access`	If the sentence refers to unequal capacity to act on climate change, emphasizing that countries, companies, or individuals do not have equal financial, technical, or institutional means to take climate action. For example, small businesses may struggle to decarbonize compared to multinational firms, and low-income households may not afford sustainable alternatives. This label reflects inequality in ability to act, not impacts or blame.
22	Intergenerational justice	`justice_intergen`	If the sentence refers to intergenerational justice, meaning that current decisions and actions should not compromise the rights or well-being of future generations. Do not apply this label to sentences that merely mention age differences or generational groups (e.g., "young people are protesting") unless the statement clearly relates to moral responsibility across generations or the long-term consequences of climate inaction.
Political Frame			
23	**Political Frame (Primary Category)**	`political_frame`	If the sentence frames climate change as a political issue, including references to policy adoption or implementation, political debates or controversies, political positioning by parties or leaders on climate issues, and public opinion data related to climate change. This label captures the role of governance, decision-making, and political dynamics in climate discourse.
24	Policy action	`pol_action`	If the sentence refers to the formal approval, adoption, or enactment of a policy, plan, or strategy by a government at the local, provincial, national, or international level. This includes the passing of new legislation, regulations, government programs, or official commitments related to climate or environmental issues (e.g., approval of a carbon pricing law, ratification of a renewable energy policy, adoption of a climate action framework).
25	Political debate	`pol_debate`	If the sentence refers to a political debate or discussion about climate change, particularly focusing on the differing views or disagreements on climate policies, the effectiveness of past actions, or the responsibility of different levels of government (e.g., federal vs. provincial). This includes disagreements about specific climate initiatives, such as proposed or enacted policies, laws, or funding allocations.
26	Political positioning	`pol_position`	If the sentence refers to the explicit stance or position of a politician or political party on climate change or climate-related policies. This includes statements where a political figure or party expresses their support or opposition to specific climate actions (e.g., carbon taxes, energy transitions, or climate mitigation strategies), or when they make electoral promises regarding climate policies.

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
27	Public opinion data	`pol_opinion`	If the sentence refers to data, surveys, polls, or studies measuring public opinion, attitudes, beliefs, or perceptions related to climate change, environmental policies, or climate action. This label captures quantitative or qualitative information on how the public views climate issues.
Scientific Frame			
28	**Scientific Frame (Primary Category)**	`scientific_frame`	If the sentence frames climate change as a scientific issue, including references to scientific research, discoveries, explanations, or assessments of climate change. This includes mentions of scientific consensus or controversy, as well as expressions of uncertainty or certainty in climate science. This label captures the role of scientific knowledge and expertise in understanding and communicating climate change.
29	Scientific debate	`sci_debate`	If the sentence refers to a past or ongoing debate or disagreement within the scientific community about aspects of climate science or climate-related technologies. This includes contested interpretations of data, disputes about methodologies, or ethical debates surrounding proposed scientific or technological solutions such as geoengineering. It does not include general questioning of climate science or statements that challenge the validity or sufficiency of climate evidence.
30	Popularisation or scientific discovery	`sci_discovery`	If the sentence explains or describes climate science mostly grounded in the natural sciences—such as chemistry, physics, or biology—or presents scientific or technological discoveries related to climate change. This includes explanations of mechanisms (e.g., the greenhouse effect, climate feedbacks), climate modeling, or innovations such as carbon capture. It does not apply to statements that merely describe environmental conditions or trends without explaining an underlying scientific process.
31	Questioning of climate science	`sci_skepticism`	If the sentence questions the validity, accuracy, or sufficiency of climate science. This includes claims that climate projections are flawed, evidence is lacking, or conclusions are exaggerated. It only applies when science is being challenged or delegitimized, not when scientists acknowledge limitations, complexity, or areas of ongoing research.
32	Defense of climate science	`sci_defense`	If the sentence defends the robustness or credibility of climate science in response to skepticism or denial. This includes assertions that reinforce the scientific consensus, affirm the validity of climate models, or rebut claims that climate science is incorrect or misleading. It only applies when climate science is defended, not when it is neutrally presented.
Environmental Frame			

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
33	**Environmental Frame (Primary Category)**	`environmental_frame`	If the sentence frames climate change as an environmental issue, including references to the degradation or disappearance of natural habitats, loss of biodiversity, and negative impacts on fauna and flora. This label captures ecological consequences of climate change and emphasizes the relationship between climate disruption and the natural environment.
34	Loss of natural environments	`env_habitat`	If the sentence refers to the long-term or irreversible disappearance or degradation of natural environments due to climate change. This includes situations where ecosystems or environmental features (e.g., glaciers, coral reefs) are predicted or expected to permanently disappear or significantly decline as a result of climate impacts, indicating that they will not recover. It does not apply to short-term, isolated, or recoverable events, such as one-time disasters (e.g., a single wildfire or a temporary drought) that might affect natural environments but do not necessarily imply permanent loss.
35	Loss of fauna and flora	`env_species`	If the sentence refers to the negative impacts of climate change on animal and plant species, including changes in habitat, shifts in migration or breeding patterns, increased mortality, loss of biodiversity, or threats to species survival. This label covers both temporary and ongoing effects on fauna and flora caused by climate change.
Cultural Frame			
36	**Cultural Frame (Primary Category)**	`cultural_frame`	If the sentence frames climate change as a cultural issue, including cultural depictions of climate change through art, literature, music, or film; challenges to hosting artistic or sports events due to climate impacts; the loss or disruption of Indigenous cultural practices; or the carbon footprint of the cultural and creative sectors. This label captures how climate change affects and is expressed through cultural life and identities.
37	Artistic representation	`cult_art`	If the sentence refers to cultural representations of climate change in various forms such as art, literature, music, film, documentaries, theater, or other creative expressions. This label captures how climate change is portrayed, interpreted, or communicated through cultural and artistic mediums.
38	Difficulty to host cultural or sports events	`cult_event_impact`	If the sentence refers to how climate change disrupts or challenges the organization of artistic, cultural, or sports events, such as cancelled festivals, concerts, or competitions due to extreme weather (e.g., storms, heatwaves, wildfires, or lack of snow for winter sports). This label captures climate-related obstacles to event planning or continuity.
39	Loss of indigenous practices	`cult_indigenous`	If the sentence refers to the erosion or disruption of traditional cultural practices among Indigenous communities due to climate change, such as changes in animal migration, plant cycles, or land access that affect food systems, ceremonies, or knowledge transmission. This label emphasizes climate-driven impacts on Indigenous cultural continuity.

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
40	Carbon footprint of the cultural and sports sectors	`cult_footprint`	If the sentence refers to the environmental or carbon footprint of the cultural, artistic, or sports sectors, including the emissions and resource use associated with activities such as touring, international travel, energy-intensive venues, unsustainable stage or set production, or event-related waste. This label covers the ecological impact of producing and hosting cultural or sports events, not the impacts of climate on those events.
PRIMARY CATEGORIES			
Actors/Messengers			
41	**Presence of Messengers (Primary Category)**	`messenger`	If the sentence includes mentions or quotes of persons or organizations by journalists, whether through direct quotations or indirect references. This label captures the presence of sources, experts, officials, activists, or other actors cited or named within the text. A single individual or organization may be associated with multiple types of expertise or roles (e.g., a doctor would fall both under Health expert and Natural scientist).
42	Health expert	`msg_health`	If the quoted or reported statement comes from a person or organization with medical or health-related recognized expertise. This includes individuals working in the fields of medicine, public health, healthcare administration, or biomedical research, whether in a professional, academic, or governmental capacity (e.g., professor of public health, doctor, nurse, family physician, public health officer, WHO representative, Red Cross medical coordinator, Doctors without Borders spokesperson, Canadian Medical Association representative, certified health consultants, minister of Health).
43	Economic expert	`msg_economic`	If the quoted or reported statement comes from a person or organization with economic or financial recognized expertise. This includes individuals working in the fields of economics, finance, business, or economic policy, whether in a professional, academic, private sector, or governmental capacity (e.g., economist, finance minister, central bank official, business executive, investment analyst, professor of economics, representative of a chamber of commerce, World Bank or IMF spokesperson, economic think tank expert, labour market analyst).
44	Security expert	`msg_security`	If the quoted or reported statement comes from a person or organization with recognized expertise in security or defence matters. This includes individuals with academic, professional, or operational experience in military, national security, intelligence, or strategic affairs (e.g., professor in defense or security studies, active or retired member of the armed forces, intelligence agency representative, conflict resolution specialist, national defence minister, security analyst, representative of a security-focused think tank or NGO).

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
45	Legal expert	`msg_legal`	If the quoted or reported statement comes from a person or organization with recognized expertise in law or legal affairs. This includes individuals working in legal professions, legal academics, the judiciary, or justice-related government roles (e.g., lawyer, prosecutor, law professor, minister of justice, attorney general, legal aid representative, bar association representative, legal consultants).
46	Cultural or Sport expert	`msg_cultural`	If the quoted or reported statement comes from a person with recognized expertise in sports, arts, or culture (e.g., Olympian, coach, professor of literature, museum director, filmmaker, writer, festival organizer, musician, artist, indigenous knowledge keeper).
47	Natural scientist	`msg_scientist`	If the quoted or reported statement comes from a person or organization with recognized expertise in the natural or hard sciences. This includes individuals working in fields such as climatology, environmental science, physics, chemistry, biology, earth sciences, engineering, or related disciplines, whether in academic, governmental, or research institutions (e.g., climate scientist, environmental researcher, oceanographer, physicist, ecologist, geologist, engineer, professor of atmospheric sciences, IPCC contributor, Environment Canada scientist, research institute spokesperson).
48	Social scientist	`msg_social`	If the quoted or reported statement comes from a person with expertise in social sciences, helping to understand how society perceives, responds to, and governs issues like climate change (e.g., sociologist, political scientist, media scholar, economist).
49	Activist	`msg_activist`	If the quoted or reported statement comes from a person or organization known primarily for advocacy or activism. This includes individuals or groups engaged in campaigning, protesting, or raising awareness on environmental, social, or climate-related issues, whether through grassroots organizing, NGOs, or international movements (e.g., climate activist, environmental campaigner, Indigenous land defender, Fridays for Future organizer, Greenpeace spokesperson, Extinction Rebellion member, citizen activist, youth climate leader, environmental justice advocate).
50	Public official	`msg_official`	If the quoted or reported statement comes from a politician, elected official, government representative, or public decision-maker at any level of government. This includes individuals holding political office or involved in policymaking, regulation, or governance, whether municipal, provincial, federal, or international (e.g., prime minister, environment minister, member of parliament, city councillor, premier, senator, government spokesperson, cabinet member, UN climate negotiator, provincial regulator, parliamentary committee chair).
Events			

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
51	**Presence of Events (Primary Category)**	`event`	If the sentence refers to the occurrence or mention of one or more events. This includes extreme meteorological events (e.g., storms, floods, heatwaves), political or governmental events (e.g., elections, policy announcements, trials, judiciary decisions), meetings or conferences, publications, cultural events, protests, or other significant happenings.
52	Extreme meteorological event	`evt_weather`	If the sentence refers to the occurrence of an extreme meteorological or climate-related event that causes or risks causing significant disruption, damage, or harm. This includes events such as heatwaves, floods, wildfires, hurricanes, droughts, storms, or other weather-related disasters, whether sudden or prolonged.
53	Meeting	`evt_meeting`	If the sentence refers to an official visit, conference, summit, forum, meeting, convention, or assembly that brings together people to discuss, negotiate, or present on specific topics at the local, national, or international levels (e.g., town hall meeting, premiers' conference, ministerial roundtable, state visit, UN conference, international forum on clean energy, G7 meeting, COP28, World Economic Forum).
54	Publication	`evt_publication`	If the sentence refers to the release or publication of a document such as a report, academic article, survey results, op-ed, policy brief, or white paper, often aimed at informing the public, guiding decision-making, or contributing to public discourse (e.g., IPCC report, Statistics Canada survey, university study on climate impacts, editorial on climate policy, NGO report on emissions, think tank publication).
55	Election	`evt_election`	If the sentence refers to the occurrence of an election at the local, provincial, national, or international level, including general elections, by-elections, leadership races, referenda, or electoral campaigns (e.g., federal election, municipal by-election, provincial leadership race, EU parliamentary elections).
56	New policy	`evt_policy`	If the sentence refers to the announcement or unveiling of a policy, plan, or strategy at the local, provincial, national, or international level, including new legislation, regulations, government programs, or official commitments, including those related to climate or environmental issues (e.g., carbon pricing plan, renewable energy strategy, climate action framework).
57	Judiciary decision	`evt_judiciary`	If the sentence refers to the occurrence of a trial, court ruling, legal proceeding, or judicial decision at any level of the judicial system, including constitutional challenges, environmental lawsuits, regulatory hearings, or Supreme Court decisions, including those related to climate or environmental issues (e.g., court ruling on carbon pricing, environmental class action lawsuit, constitutional challenge to climate legislation, tribunal decision on pipeline approval).

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
58	Cultural or Sports event	`evt_cultural`	If the sentence refers to the organization or hosting of a sports, artistic, or cultural event (e.g., Olympics games, national hockey tournament, local marathon, film screening, music concert, theatre performance, book launch, mural festival, photography exhibit).
59	Protest	`evt_protest`	If the sentence refers to the organization of a protest (e.g., climate strike, anti-pipeline protest, extinction rebellion action, anti-racism march, women's right rally, union-led protest for better wages).
Solutions			
60	**Presence of Solutions (Primary Category)**	`solution`	If the sentence refers to solutions addressing climate change, including mitigation efforts to reduce greenhouse gas emissions or adaptation strategies to manage climate impacts. This includes direct mentions of specific policies, technologies, behaviors, or initiatives, as well as indirect references to approaches aimed at combating or coping with climate change.
61	Mitigation strategy	`sol_mitigation`	If the sentence refers to mitigation solutions designed to reduce or prevent the emission of greenhouse gases and limit the extent of future climate change. These strategies aim to address the root causes of climate change by lowering carbon footprints, enhancing carbon sinks, or transitioning to low-carbon technologies (e.g., investing in renewable energy, improving energy efficiency, implementing carbon pricing, promoting public transit, reforestation, phasing out fossil fuels).
62	Adaptation strategy	`sol_adaptation`	If the sentence refers to adaptation solutions designed to reduce vulnerability to the negative effects of climate change by adjusting or preparing for those impacts. These strategies involve actions to manage the risks and enhance resilience, particularly by reducing exposure to climate-related hazards and minimizing harm to people, infrastructure, and ecosystems (e.g., providing cooling centres during heat waves, restoring wetlands to limit flooding, altering farming techniques to cope with changing weather patterns).
EMOTIONAL TONE			
63	Emotion: positive	`tone_positive`	If the tone is generally optimistic, reassuring, or enthusiastic, highlighting favorable aspects or expressing confidence in climate solutions or actions.
64	Emotion: negative	`tone_negative`	If the tone is generally alarming, critical, or pessimistic, especially if the sentence expresses concern, fear, or frustration about climate change or its impacts.
65	Emotion: neutral	`tone_neutral`	If the sentence is mainly informative or descriptive without an evaluative tone, presenting facts or balanced perspectives without strong emotional content.
GEOGRAPHIC FOCUS			

Continued on next page

Supplementary Table S3 – *Continued from previous page*

#	Category	Code	Description
66	Mention of Canada	`canada`	If the sentence refers to Canada in any context, domestic or international, including but not limited to climate change impacts on Canadian territory, as well as Canada's policies, actions, or roles.
URGENCY/ALARMISM			
67	Urgency to act	`urgency`	If the sentence conveys a strong sense of urgency or alarmism, emphasizing the immediate need for action or highlighting critical risks and dangers related to climate change, often with a tone that signals warning or crisis.

* Categories marked with an asterisk were included in the annotation protocol but had insufficient positive training examples. This table enumerates the 67 *candidate* categories of the framework; the deployed database carries 65 of them. Two candidates (`health_pos_impact`, `health_footprint`) had no positive examples in either language and were excluded entirely. Two deployed categories (`security_military`, `security_disruption`) could not be trained in English and are annotated on French sentences only (their English annotations are `NULL` by construction). The deployed model suite therefore counts $65 \times 2 - 2 = 128$ models. Because this table numbers the 67 candidates while `CCF_reliability_tiers` numbers the 65 deployed categories, the two numbering schemes differ from the Health frame onward: always match categories on the `code` column, never on the row number.

Supplementary Table S4. Complete model training performance metrics for all annotation categories.

#	Category	Code	F1 (Class 1)		F1 (Class 0)		Macro F1	
			EN	FR	EN	FR	EN	FR
MAIN FRAMES								
Economic Frame								
1	**Economic Frame (Primary Category)**	`economic_frame`	0.745	0.814	0.944	0.957	0.845	0.885
2	Negative impacts of climate change on the economy	`eco_neg_impact`	0.727	0.889	0.914	0.944	0.821	0.917
3	Positive impacts of climate change on the economy	`eco_pos_impact`	0.333	0.400	0.995	0.996	0.664	0.698
4	Economic disadvantages of climate action	`eco_cost`	0.615	0.500	0.848	0.913	0.732	0.707
5	Economic benefits of climate action	`eco_benefit`	0.500	0.516	0.952	0.981	0.726	0.749
6	Carbon footprint of the economic sector	`eco_footprint`	0.857	0.857	0.938	0.950	0.897	0.904
Health Frame								
7	**Health Frame (Primary Category)**	`health_frame`	0.800	0.667	0.989	0.994	0.894	0.830
8	Negative impacts of climate change on health	`health_neg_impact`	0.909	0.857	0.000	0.000	0.455	0.429
9	Health co-benefits of climate action	`health_cobenefit`	0.400	0.011	0.571	0.227	0.486	0.119
Security Frame								
10	**Security Frame (Primary Category)**	`security_frame`	0.870	0.800	0.996	0.997	0.933	0.898
11	Presence of climate refugees	`security_refugees`	1.000	1.000	1.000	1.000	1.000	1.000

Continued on next page

Supplementary Table S4 – Continued from previous page

#	Category	Code	F1 (Class 1)		F1 (Class 0)		Macro F1	
			EN	**FR**	**EN**	**FR**	**EN**	**FR**
12	Conflict	`security_conflict`	1.000	1.000	1.000	1.000	1.000	1.000
13	Post-disaster military assistance	`security_military`	–	0.010	–	0.043	–	0.026
14	Disruption of military operations	`security_disruption`	–	1.000	–	1.000	–	1.000
Justice Frame								
15	**Justice Frame (Primary Category)**	`justice_frame`	0.719	0.717	0.975	0.981	0.847	0.849
16	Winners and losers of climate action	`justice_winners`	0.667	0.667	0.933	0.923	0.800	0.795
17	Differentiated responsibility	`justice_responsibility`	0.857	0.750	0.909	0.750	0.883	0.750
18	Unequal vulnerability to climate change	`justice_vulnerability`	0.571	0.600	0.992	0.995	0.782	0.798
19	Unequal access to climate action	`justice_access`	0.364	0.625	0.000	0.993	0.182	0.809
20	Intergenerational justice	`justice_intergen`	0.933	0.800	0.999	0.999	0.966	0.899
Political Frame								
21	**Political Frame (Primary Category)**	`political_frame`	0.808	0.774	0.897	0.888	0.853	0.831
22	Policy action	`pol_action`	0.621	0.667	0.985	0.966	0.803	0.816
23	Political debate	`pol_debate`	0.916	0.966	0.222	0.571	0.569	0.768
24	Political positioning	`pol_position`	0.750	0.800	0.946	0.989	0.848	0.895
25	Public opinion data	`pol_opinion`	0.909	1.000	0.999	1.000	0.954	1.000
Scientific Frame								
26	**Scientific Frame (Primary Category)**	`scientific_frame`	0.784	0.702	0.953	0.962	0.869	0.832
27	Scientific debate	`sci_debate`	0.737	0.500	0.848	0.875	0.793	0.688

Continued on next page

Supplementary Table S4 – Continued from previous page

#	Category	Code	F1 (Class 1)		F1 (Class 0)		Macro F1	
			EN	**FR**	**EN**	**FR**	**EN**	**FR**
28	Popularisation or scientific discovery	`sci_discovery`	0.828	0.938	0.800	0.750	0.814	0.844
29	Questioning of climate science	`sci_skepticism`	0.727	0.333	0.927	0.995	0.827	0.664
30	Defense of climate science	`sci_defense`	0.571	0.000	0.933	0.919	0.752	0.459
Environmental Frame								
31	**Environmental Frame (Primary Category)**	`environmental_frame`	0.842	0.625	0.989	0.980	0.915	0.802
32	Loss of natural environments	`env_habitat`	1.000	0.857	1.000	0.800	1.000	0.829
33	Loss of fauna and flora	`env_species`	0.889	0.857	0.857	0.800	0.873	0.829
Cultural Frame								
34	**Cultural Frame (Primary Category)**	`cultural_frame`	0.773	0.833	0.986	0.993	0.879	0.913
35	Artistic representation	`cult_art`	1.000	1.000	1.000	1.000	1.000	1.000
36	Difficulty to host cultural or sports events	`cult_event_impact`	0.706	1.000	0.993	1.000	0.850	1.000
37	Loss of indigenous practices	`cult_indigenous`	1.000	0.026	1.000	0.899	1.000	0.462
38	Carbon footprint of the cultural and sports sectors	`cult_footprint`	1.000	0.005	1.000	0.143	1.000	0.074
PRIMARY CATEGORIES								
Actors/Messengers								
39	**Presence of Messengers (Primary Category)**	`messenger`	0.912	0.929	0.904	0.915	0.908	0.922
40	Health expert	`msg_health`	0.857	0.909	0.997	0.999	0.927	0.954
41	Economic expert	`msg_economic`	0.600	0.750	0.970	0.973	0.785	0.861
42	Security expert	`msg_security`	0.667	1.000	0.993	1.000	0.830	1.000

Continued on next page

Supplementary Table S4 – Continued from previous page

#	Category	Code	F1 (Class 1)		F1 (Class 0)		Macro F1	
			EN	FR	EN	FR	EN	FR
43	Legal expert	msg_legal	0.667	0.769	0.999	0.996	0.833	0.883
44	Cultural or Sport expert	msg_cultural	1.000	0.556	1.000	0.990	1.000	0.773
45	Natural scientist	msg_scientist	0.789	0.833	0.977	0.971	0.883	0.902
46	Social scientist	msg_social	0.769	0.645	0.996	0.986	0.883	0.816
47	Activist	msg_activist	0.571	1.000	0.978	1.000	0.775	1.000
48	Public official	msg_official	0.703	0.923	0.897	0.964	0.800	0.943
Events								
49	**Presence of Events (Primary Category)**	event	0.794	0.819	0.935	0.932	0.865	0.876
50	Extreme meteorological event	evt_weather	1.000	0.857	1.000	0.968	1.000	0.912
51	Meeting	evt_meeting	0.824	0.957	0.936	0.982	0.880	0.969
52	Publication	evt_publication	0.873	1.000	0.990	1.000	0.931	1.000
53	Election	evt_election	1.000	0.800	1.000	0.998	1.000	0.899
54	New policy	evt_policy	0.769	0.727	0.943	0.954	0.856	0.841
55	Judiciary decision	evt_judiciary	1.000	1.000	1.000	1.000	1.000	1.000
56	Cultural or Sports event	evt_cultural	0.333	1.000	0.995	1.000	0.664	1.000
57	Protest	evt_protest	0.889	0.727	0.999	0.996	0.944	0.862
Solutions								
58	**Presence of Solutions (Primary Category)**	solution	0.737	0.878	0.914	0.944	0.825	0.911
59	Mitigation strategy	sol_mitigation	0.750	0.812	0.935	0.942	0.842	0.877
60	Adaptation strategy	sol_adaptation	0.696	0.800	0.991	0.989	0.843	0.894
EMOTIONAL TONE								
61	Emotion: positive	tone_positive	0.526	0.690	0.966	0.968	0.746	0.829
62	Emotion: negative	tone_negative	0.706	0.765	0.785	0.843	0.745	0.804
63	Emotion: neutral	tone_neutral	0.741	0.789	0.676	0.693	0.709	0.741
GEOGRAPHIC FOCUS								

Continued on next page

Supplementary Table S4 – Continued from previous page

#	Category	Code	F1 (Class 1)		F1 (Class 0)		Macro F1	
			EN	FR	EN	FR	EN	FR
64	Mention of Canada	`canada`	0.942	0.964	0.968	0.980	0.955	0.972
URGENCY TO ACT								
65	Urgency to act	`urgency`	0.591	0.649	0.975	0.984	0.783	0.816
AVERAGE PERFORMANCE METRICS								
		English Average	0.769	–	0.905	–	0.837	–
		French Average	–	0.737	–	0.894	–	0.816
		Overall Average	0.753		0.900		0.826	
TOTAL: 65 ANNOTATION CATEGORIES *(65 deployed categories; 63 with models in both languages, 2 with a French model only*)*								

Note. F1 scores at the best epoch retained by the combined-criterion best-model selector ($0.98 \cdot F_1^{(1)} + 0.02 \cdot$ macro-F_1). EN columns report metrics for the English BERT-base-cased classifiers; FR columns report metrics for the French CamemBERT-base classifiers. A "–" in one language column means the category could not be trained in that language for lack of positive examples: this concerns exactly two English models (`security_military`, `security_disruption`), which is why the deployed suite counts $65 \times 2 - 2 = 128$ models; the two candidate categories excluded in both languages do not appear in this table (see the note to Supplementary Table S3). The dataset distribution behind each row is reported in Supplementary Table S5.

Supplementary Table S5. Training and validation dataset distribution across all annotation categories.

#	Category	Code	English				French			
			Training		Validation		Training		Validation	
			Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
MAIN FRAMES										
Economic Frame										
1	**Economic Frame (Primary Category)**	`economic_frame`	218	1067	24	118	252	1164	28	129
2	Negative impacts of climate change on the economy	`eco_neg_impact`	61	158	6	17	75	178	8	19
3	Positive impacts of climate change on the economy	`eco_pos_impact`	14	205	1	22	7	245	1	27
4	Economic disadvantages of climate action	`eco_cost`	63	156	6	17	42	211	4	23
5	Economic benefits of climate action	`eco_benefit`	36	183	3	20	48	205	5	22
6	Carbon footprint of the economic sector	`eco_footprint`	67	152	7	16	77	176	8	19
Health Frame										
7	**Health Frame (Primary Category)**	`health_frame`	57	1228	6	136	37	1379	4	153
8	Negative impacts of climate change on health	`health_neg_impact`	47	10	5	1	32	5	3	1
9	Health co-benefits of climate action	`health_cobenefit`	8	49	1	5	4	33	1	3
Security Frame										
10	**Security Frame (Primary Category)**	`security_frame`	19	1266	2	140	19	1397	2	155

Continued on next page

Supplementary Table S5 – Continued from previous page

#	Category	Code	English				French			
			Training		Validation		Training		Validation	
			Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
11	Presence of climate refugees	security_refugees	10	9	1	1	7	12	1	1
12	Conflict	security_conflict	6	13	1	1	7	12	1	1
13	Post-disaster military assistance	security_military	0	18	1	2	4	15	1	1
14	Disruption of military operations	security_disruption	0	18	1	2	1	18	1	1
Justice Frame										
15	**Justice Frame (Primary Category)**	justice_frame	90	1196	9	132	81	1335	9	148
16	Winners and losers of climate action	justice_winners	20	70	2	7	20	62	2	6
17	Differentiated responsibility	justice_responsibility	38	52	4	5	41	41	4	4
18	Unequal vulnerability to climate change	justice_vulnerability	8	81	1	9	9	72	1	8
19	Unequal access to climate action	justice_access	21	69	2	7	20	62	2	6
20	Intergenerational justice	justice_intergen	14	76	1	8	6	75	1	8
Political Frame										
21	**Political Frame (Primary Category)**	political_frame	416	869	46	96	440	977	48	108
22	Policy action	pol_action	62	355	6	39	58	382	6	42
23	Political debate	pol_debate	344	72	38	8	396	45	43	4
24	Political positioning	pol_position	61	356	6	39	35	405	3	45
25	Public opinion data	pol_opinion	21	396	2	43	20	420	2	46
Scientific Frame										

Continued on next page

Supplementary Table S5 – Continued from previous page

#	Category	Code	English				French			
			Training		Validation		Training		Validation	
			Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
26	**Scientific Frame (Primary Category)**	`scientific_frame`	243	1042	27	115	185	1232	20	136
27	Scientific debate	`sci_debate`	98	146	10	16	46	139	5	15
28	Popularisation or scientific discovery	`sci_discovery`	126	117	14	13	144	41	16	4
29	Questioning of climate science	`sci_skepticism`	50	194	5	21	17	169	1	18
30	Defense of climate science	`sci_defense`	37	207	4	22	32	153	3	17
Environmental Frame										
31	**Environmental Frame (Primary Category)**	`environmental_frame`	84	1201	9	133	63	1354	6	150
32	Loss of natural environments	`env_habitat`	50	35	5	3	35	28	3	3
33	Loss of fauna and flora	`env_species`	42	43	4	4	32	31	3	3
Cultural Frame										
34	**Cultural Frame (Primary Category)**	`cultural_frame`	44	1242	4	137	45	1371	5	152
35	Artistic representation	`cult_art`	20	24	2	2	18	27	2	3
36	Difficulty to host cultural or sports events	`cult_event_impact`	11	33	1	3	21	25	2	2
37	Loss of indigenous practices	`cult_indigenous`	10	34	1	3	4	41	1	4
38	Carbon footprint of the cultural and sports sectors	`cult_footprint`	1	42	1	4	2	43	1	4
PRIMARY CATEGORIES										
Actors/Messengers										

Continued on next page

Supplementary Table S5 – Continued from previous page

#	Category	Code	English				French			
			Training		Validation		Training		Validation	
			Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
39	**Presence of Messengers (Primary Category)**	`messenger`	657	629	72	69	736	681	81	75
40	Health expert	`msg_health`	10	647	1	71	8	728	1	80
41	Economic expert	`msg_economic`	55	602	6	66	67	669	7	74
42	Security expert	`msg_security`	3	653	1	72	10	726	1	80
43	Legal expert	`msg_legal`	1	655	1	72	4	731	1	81
44	Cultural or Sport expert	`msg_cultural`	13	644	1	71	9	727	1	80
45	Natural scientist	`msg_scientist`	107	550	11	61	117	620	12	68
46	Social scientist	`msg_social`	9	648	1	71	26	711	2	78
47	Activist	`msg_activist`	53	604	5	67	55	681	6	75
48	Public official	`msg_official`	171	486	18	54	223	513	24	57
Events										
49	**Presence of Events (Primary Category)**	`event`	298	987	33	109	355	1062	39	117
50	Extreme meteorological event	`evt_weather`	72	226	8	25	59	297	6	32
51	Meeting	`evt_meeting`	75	224	8	24	100	255	11	28
52	Publication	`evt_publication`	74	225	8	24	114	242	12	26
53	Election	`evt_election`	16	283	1	31	18	337	2	37
54	New policy	`evt_policy`	55	243	6	27	60	296	6	32
55	Judiciary decision	`evt_judiciary`	10	288	1	32	10	345	1	38
56	Cultural or Sports event	`evt_cultural`	2	296	1	32	11	344	1	38
57	Protest	`evt_protest`	11	288	1	31	8	347	1	38
Solutions										
58	**Presence of Solutions (Primary Category)**	`solution`	314	972	34	107	415	1001	46	111
59	Mitigation strategy	`sol_mitigation`	279	35	31	3	360	55	40	6

Continued on next page

Supplementary Table S5 – Continued from previous page

#	Category	Code	English				French			
			Training		Validation		Training		Validation	
			Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
60	Adaptation strategy	`sol_adaptation`	46	268	5	29	64	351	7	39
EMOTIONAL TONE										
61	Emotion: positive	`tone_positive`	103	1182	11	131	151	1266	16	140
62	Emotion: negative	`tone_negative`	491	794	54	88	537	880	59	97
63	Emotion: neutral	`tone_neutral`	686	599	76	66	726	691	80	76
GEOGRAPHIC FOCUS										
64	Mention of Canada	`canada`	445	840	49	93	493	924	54	102
URGENCY TO ACT										
65	Urgency to act	`urgency`	66	1219	7	135	63	1354	6	150

Note. Positive (**Pos**) and negative (**Neg**) sentence counts in the training and validation splits, by language. Categories with zero counts in every split could not be trained due to insufficient positive examples; the resulting per-category training metrics are reported in Supplementary Table S4.

Supplementary Table S6. Named Entity Recognition performance.

Table S11: Named Entity Recognition (`ner_entities` variable) model performance metrics

Language	**Model**	**PER F1**	**ORG F1**	**LOC F1**
English	BERT-base-NER	0.961	0.811	0.925
French	spaCy fr_core_news_lg	0.880	–	–
	CamemBERT-NER	–	0.824	0.929

Note: The hybrid approach for French combines spaCy for person entities (PER) with CamemBERT-NER for organization (ORG) and location (LOC) entities based on empirical evaluation on the dataset. Performance metrics are from the original model documentation.

Supplementary Table S7. Detailed validation performance metrics on the 1,000-sentence gold standard.

#	Category	Code	F1 Macro			F1 Micro			F1 Weighted			Support		
			EN	FR	ALL	EN	FR	ALL	EN	FR	ALL	EN	FR	ALL
MAIN FRAMES														
Economic Frame														
1	**Economic Frame (Primary Category)**	`economic_frame`	0.847	0.814	0.830	0.892	0.864	0.878	0.890	0.862	0.875	120	129	249
2	Negative impacts of climate change on the economy	`eco_neg_impact`	0.820	0.741	0.783	0.848	0.810	0.828	0.856	0.833	0.842	26	18	44
3	Positive impacts of climate change on the economy	`eco_pos_impact`	0.806	0.508	0.684	0.905	0.845	0.873	0.913	0.908	0.900	12	1	13
4	Economic disadvantages of climate action	`eco_cost`	0.741	0.702	0.723	0.800	0.802	0.801	0.813	0.806	0.810	22	23	45
5	Economic benefits of climate action	`eco_benefit`	0.885	0.817	0.848	0.924	0.871	0.896	0.928	0.872	0.899	19	26	45
6	Carbon footprint of the economic sector	`eco_footprint`	0.785	0.788	0.787	0.838	0.845	0.842	0.831	0.843	0.837	30	29	59
Health Frame														
7	**Health Frame (Primary Category)**	`health_frame`	0.800	0.817	0.808	0.947	0.961	0.954	0.947	0.964	0.956	35	23	58
8	Negative impacts of climate change on health	`health_neg_impact`	0.386	0.340	0.364	0.629	0.514	0.571	0.485	0.349	0.416	22	18	40
9	Health co-benefits of climate action	`health_cobenefit`	0.252	0.103	0.186	0.257	0.114	0.186	0.301	0.023	0.177	4	4	8
Security Frame														
10	**Security Frame (Primary Category)**	`security_frame`	0.825	0.730	0.782	0.951	0.939	0.945	0.955	0.950	0.952	30	19	49
11	Presence of climate refugees	`security_refugees`	0.476	0.417	0.481	0.545	0.429	0.488	0.468	0.476	0.459	21	6	27
12	Conflict	`security_conflict`	0.358	0.286	0.324	0.364	0.286	0.326	0.398	0.286	0.346	7	6	13
13	Post-disaster military assistance	`security_military`	0.000	0.045	0.045	0.000	0.048	0.048	0.000	0.004	0.004	0	2	2
14	Disruption of military operations	`security_disruption`	0.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0	2	2
Justice Frame														
15	**Justice Frame (Primary Category)**	`justice_frame`	0.832	0.880	0.857	0.917	0.937	0.927	0.921	0.939	0.930	62	74	136

Continued on next page

Supplementary Table S7 – Continued from previous page

#	Category	Code	F1 Macro			F1 Micro			F1 Weighted			Support		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**	**EN**	**FR**	**ALL**	**EN**	**FR**	**ALL**
16	Winners and losers of climate action	`justice_winners`	0.654	0.737	0.708	0.765	0.762	0.764	0.802	0.775	0.786	10	22	32
17	Differentiated responsibility	`justice_responsibility`	0.900	0.823	0.860	0.926	0.833	0.879	0.927	0.838	0.883	19	27	46
18	Unequal vulnerability to climate change	`justice_vulnerability`	0.786	0.894	0.845	0.852	0.917	0.885	0.858	0.918	0.888	16	21	37
19	Unequal access to climate action	`justice_access`	0.475	0.796	0.651	0.481	0.833	0.661	0.456	0.835	0.674	27	23	50
20	Intergenerational justice	`justice_intergen`	0.924	0.952	0.942	0.963	0.964	0.964	0.965	0.965	0.965	10	19	29
Political Frame														
21	**Political Frame (Primary Category)**	`political_frame`	0.812	0.842	0.829	0.827	0.844	0.836	0.829	0.845	0.837	167	217	384
22	Policy action	`pol_action`	0.802	0.722	0.756	0.914	0.864	0.886	0.909	0.854	0.879	26	38	64
23	Political debate	`pol_debate`	0.469	0.650	0.564	0.694	0.820	0.763	0.595	0.781	0.699	127	175	302
24	Political positioning	`pol_position`	0.752	0.808	0.782	0.801	0.912	0.862	0.821	0.920	0.877	35	24	59
25	Public opinion data	`pol_opinion`	0.877	0.910	0.894	0.962	0.974	0.969	0.966	0.976	0.971	12	15	27
Scientific Frame														
26	**Scientific Frame (Primary Category)**	`scientific_frame`	0.891	0.890	0.890	0.951	0.949	0.950	0.948	0.947	0.947	74	75	149
27	Scientific debate	`sci_debate`	0.962	0.867	0.911	0.962	0.869	0.912	0.962	0.869	0.912	26	26	52
28	Popularisation or scientific discovery	`sci_discovery`	0.884	0.797	0.838	0.885	0.803	0.841	0.886	0.803	0.842	31	36	67
29	Questioning of climate science	`sci_skepticism`	0.783	0.768	0.775	0.846	0.836	0.841	0.860	0.857	0.858	9	9	18
30	Defense of climate science	`sci_defense`	0.769	0.460	0.682	0.846	0.852	0.850	0.862	0.785	0.843	8	9	17
Environmental Frame														
31	**Environmental Frame (Primary Category)**	`environmental_frame`	0.866	0.875	0.871	0.967	0.967	0.967	0.969	0.968	0.969	28	32	60
32	Loss of natural environments	`env_habitat`	0.804	0.573	0.721	0.806	0.707	0.753	0.802	0.637	0.730	16	26	42
33	Loss of fauna and flora	`env_species`	0.714	0.692	0.703	0.722	0.707	0.714	0.717	0.694	0.705	19	21	40
Cultural Frame														
34	**Cultural Frame (Primary Category)**	`cultural_frame`	0.802	0.581	0.708	0.923	0.880	0.901	0.931	0.920	0.921	41	11	52
35	Artistic representation	`cult_art`	0.834	0.469	0.659	0.851	0.559	0.704	0.858	0.649	0.739	18	6	24
36	Difficulty to host cultural or sports events	`cult_event_impact`	0.631	0.354	0.484	0.731	0.456	0.593	0.777	0.595	0.686	8	2	10
37	Loss of indigenous practices	`cult_indigenous`	0.824	0.415	0.595	0.910	0.618	0.763	0.922	0.749	0.825	7	1	8

Continued on next page

Supplementary Table S7 – Continued from previous page

#	Category	Code	F1 Macro			F1 Micro			F1 Weighted			Support		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**	**EN**	**FR**	**ALL**	**EN**	**FR**	**ALL**
38	Carbon footprint of the cultural and sports sectors	`cult_footprint`	0.702	0.014	0.351	0.925	0.015	0.467	0.945	0.000	0.613	2	1	3
PRIMARY CATEGORIES														
Actors/Messengers														
39	**Presence of Messengers (Primary Category)**	`messenger`	0.955	0.967	0.961	0.957	0.969	0.963	0.957	0.969	0.963	301	312	613
40	Health expert	`msg_health`	0.862	0.951	0.907	0.964	0.987	0.975	0.966	0.987	0.976	19	21	40
41	Economic expert	`msg_economic`	0.847	0.895	0.874	0.954	0.961	0.957	0.949	0.963	0.957	30	28	58
42	Security expert	`msg_security`	0.728	0.813	0.787	0.970	0.958	0.964	0.973	0.962	0.968	7	14	21
43	Legal expert	`msg_legal`	0.861	0.927	0.914	0.990	0.980	0.985	0.989	0.982	0.986	7	19	26
44	Cultural or Sport expert	`msg_cultural`	0.853	0.825	0.843	0.954	0.967	0.961	0.954	0.969	0.961	26	13	39
45	Natural scientist	`msg_scientist`	0.920	0.896	0.907	0.960	0.941	0.951	0.960	0.944	0.952	44	46	90
46	Social scientist	`msg_social`	0.938	0.948	0.944	0.987	0.984	0.985	0.987	0.984	0.986	17	24	41
47	Activist	`msg_activist`	0.934	0.903	0.918	0.977	0.964	0.970	0.976	0.963	0.970	31	33	64
48	Public official	`msg_official`	0.927	0.930	0.929	0.934	0.935	0.934	0.934	0.935	0.934	104	113	217
Events														
49	**Presence of Events (Primary Category)**	`event`	0.881	0.912	0.897	0.886	0.913	0.900	0.885	0.913	0.900	201	222	423
50	Extreme meteorological event	`evt_weather`	0.887	0.891	0.889	0.935	0.955	0.946	0.938	0.958	0.949	29	21	50
51	Meeting	`evt_meeting`	0.855	0.891	0.877	0.908	0.914	0.911	0.914	0.918	0.916	30	50	80
52	Publication	`evt_publication`	0.950	0.962	0.958	0.973	0.968	0.970	0.973	0.969	0.971	30	62	92
53	Election	`evt_election`	0.911	0.895	0.903	0.962	0.964	0.963	0.960	0.963	0.962	25	22	47
54	New policy	`evt_policy`	0.821	0.837	0.829	0.886	0.914	0.901	0.895	0.915	0.906	29	33	62
55	Judiciary decision	`evt_judiciary`	0.628	0.939	0.808	0.903	0.982	0.946	0.927	0.981	0.951	6	19	25
56	Cultural or Sports event	`evt_cultural`	0.906	0.857	0.881	0.973	0.964	0.968	0.971	0.967	0.968	17	12	29
57	Protest	`evt_protest`	0.982	0.909	0.943	0.995	0.973	0.983	0.995	0.973	0.983	16	18	34
Solutions														
58	**Presence of Solutions (Primary Category)**	`solution`	0.891	0.909	0.900	0.921	0.931	0.926	0.922	0.930	0.926	114	133	247
59	Mitigation strategy	`sol_mitigation`	0.872	0.801	0.839	0.884	0.839	0.861	0.882	0.850	0.864	76	97	173
60	Adaptation strategy	`sol_adaptation`	0.889	0.903	0.895	0.934	0.952	0.943	0.934	0.954	0.944	22	16	38
EMOTIONAL TONE														
61	Emotion: positive	`tone_positive`	0.712	0.692	0.702	0.902	0.892	0.897	0.910	0.904	0.907	38	37	75
62	Emotion: negative	`tone_negative`	0.793	0.804	0.799	0.807	0.813	0.810	0.814	0.819	0.816	151	165	316
63	Emotion: neutral	`tone_neutral`	0.756	0.798	0.777	0.764	0.809	0.787	0.765	0.807	0.787	295	306	601

Continued on next page

Supplementary Table S7 – Continued from previous page

#	Category	Code	F1 Macro			F1 Micro			F1 Weighted			Support		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**	**EN**	**FR**	**ALL**	**EN**	**FR**	**ALL**
GEOGRAPHIC FOCUS														
64	Mention of Canada	`canada`	0.986	0.978	0.982	0.986	0.978	0.982	0.986	0.978	0.982	235	255	490
URGENCY TO ACT														
65	Urgency to act	`urgency`	0.874	0.835	0.853	0.976	0.965	0.970	0.977	0.967	0.972	23	25	48
OVERALL PERFORMANCE														
		All Categories	**0.869**	**0.864**	**0.866**	**0.909**	**0.902**	**0.905**	**0.911**	**0.905**	**0.908**	**3069**	**3332**	**6401**

Note. F1 macro, F1 micro, F1 weighted, and positive-class support computed on the 1,000-sentence gold standard. EN, FR, and ALL columns report metrics on the English subset, the French subset, and their union. **OVERALL PERFORMANCE** in the last row reports the *pooled* aggregates—all per-category predictions are pooled into a single confusion matrix before the scores are computed—and is therefore not the mean of the per-category rows. The unweighted mean of the 65 per-category macro F1 values (ALL column) is 0.773; this is the quantity compared like-for-like with the training-phase mean of 0.826 in the main text (Technical Validation). Support columns report the number of expert-confirmed positive examples; 23 of the 65 categories fall below the sampling target of 40 (see the main text on how to weigh low-support estimates). Inter-coder agreement statistics on the same 1,000 sentences are reported in Supplementary Tables S9 and S10.

Supplementary Table S8. Database-wide distribution of annotation categories across the 9.9 million sentences of CCF_processed_data.

#	Category	Code	Count			Proportion (%)		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**
MAIN FRAMES								
Economic Frame								
1	**Economic Frame (Primary Category)**	`economic_frame`	1,312,306	206,394	1,518,700	15.88	12.56	15.33
2	Negative impacts of climate change on the economy	`eco_neg_impact`	208,232	35,469	243,701	2.52	2.16	2.46
3	Positive impacts of climate change on the economy	`eco_pos_impact`	1,216	2,198	3,414	0.01	0.13	0.03
4	Economic disadvantages of climate action	`eco_cost`	549,771	45,273	595,044	6.65	2.75	6.01
5	Economic benefits of climate action	`eco_benefit`	79,448	32,961	112,409	0.96	2.01	1.13
6	Carbon footprint of the economic sector	`eco_footprint`	308,856	67,727	376,583	3.74	4.12	3.80
Health Frame								
7	**Health Frame (Primary Category)**	`health_frame`	116,449	24,852	141,301	1.41	1.51	1.43
8	Negative impacts of climate change on health	`health_neg_impact`	116,449	24,852	141,301	1.41	1.51	1.43
9	Health co-benefits of climate action	`health_cobenefit`	68,998	23,026	92,024	0.83	1.40	0.93
Security Frame								
10	**Security Frame (Primary Category)**	`security_frame`	34,882	11,402	46,284	0.42	0.69	0.47

Continued on next page

Supplementary Table S8 – Continued from previous page

#	Category	Code	Count			Proportion (%)		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**
11	Presence of climate refugees	`security_refugees`	22,723	4,556	27,279	0.27	0.28	0.28
12	Conflict	`security_conflict`	21,758	9,969	31,727	0.26	0.61	0.32
13	Post-disaster military assistance	`security_military`	0	11,400	11,400	0.00	0.69	0.12
14	Disruption of military operations	`security_disruption`	0	4	4	0.00	0.00	0.00
Justice Frame								
15	**Justice Frame (Primary Category)**	`justice_frame`	290,014	22,364	312,378	3.51	1.36	3.15
16	Winners and losers of climate action	`justice_winners`	47,725	9,625	57,350	0.58	0.59	0.58
17	Differentiated responsibility	`justice_responsibility`	61,525	9,362	70,887	0.74	0.57	0.72
18	Unequal vulnerability to climate change	`justice_vulnerability`	13,845	2,423	16,268	0.17	0.15	0.16
19	Unequal access to climate action	`justice_access`	225,960	5,067	231,027	2.73	0.31	2.33
20	Intergenerational justice	`justice_intergen`	28,753	1,295	30,048	0.35	0.08	0.30
Political Frame								
21	**Political Frame (Primary Category)**	`political_frame`	2,548,832	447,833	2,996,665	30.84	27.25	30.24
22	Policy action	`pol_action`	219,650	21,051	240,701	2.66	1.28	2.43
23	Political debate	`pol_debate`	2,494,072	438,459	2,932,531	30.17	26.68	29.60
24	Political positioning	`pol_position`	901,142	52,613	953,755	10.90	3.20	9.63
25	Public opinion data	`pol_opinion`	48,852	8,857	57,709	0.59	0.54	0.58
Scientific Frame								

Continued on next page

Supplementary Table S8 – Continued from previous page

#	Category	Code	Count			Proportion (%)		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**
26	**Scientific Frame (Primary Category)**	`scientific_frame`	545,323	86,067	631,390	6.60	5.24	6.37
27	Scientific debate	`sci_debate`	221,198	15,847	237,045	2.68	0.96	2.39
28	Popularisation or scientific discovery	`sci_discovery`	366,775	72,724	439,499	4.44	4.43	4.44
29	Questioning of climate science	`sci_skepticism`	67,259	2,250	69,509	0.81	0.14	0.70
30	Defense of climate science	`sci_defense`	50,380	0	50,380	0.61	0.00	0.51
Environmental Frame								
31	**Environmental Frame (Primary Category)**	`environmental_frame`	218,374	88,446	306,820	2.64	5.38	3.10
32	Loss of natural environments	`env_habitat`	113,064	74,042	187,106	1.37	4.51	1.89
33	Loss of fauna and flora	`env_species`	126,270	61,208	187,478	1.53	3.72	1.89
Cultural Frame								
34	**Cultural Frame (Primary Category)**	`cultural_frame`	306,791	96,179	402,970	3.71	5.85	4.07
35	Artistic representation	`cult_art`	212,176	49,753	261,929	2.57	3.03	2.64
36	Difficulty to host cultural or sports events	`cult_event_impact`	25,402	42,776	68,178	0.31	2.60	0.69
37	Loss of indigenous practices	`cult_indigenous`	7,890	21,917	29,807	0.10	1.33	0.30
38	Carbon footprint of the cultural and sports sectors	`cult_footprint`	13	96,068	96,081	0.00	5.85	0.97
PRIMARY CATEGORIES								
Actors/Messengers								
39	**Presence of Messengers (Primary Category)**	`messenger`	4,131,733	744,674	4,876,407	49.99	45.31	49.21

Continued on next page

Supplementary Table S8 – Continued from previous page

#	Category	Code	Count			Proportion (%)		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**
40	Health expert	`msg_health`	51,198	9,371	60,569	0.62	0.57	0.61
41	Economic expert	`msg_economic`	229,368	84,782	314,150	2.78	5.16	3.17
42	Security expert	`msg_security`	19	11,079	11,098	0.00	0.67	0.11
43	Legal expert	`msg_legal`	7	5,968	5,975	0.00	0.36	0.06
44	Cultural or Sport expert	`msg_cultural`	130,523	20,405	150,928	1.58	1.24	1.52
45	Natural scientist	`msg_scientist`	299,413	84,448	383,861	3.62	5.14	3.87
46	Social scientist	`msg_social`	34,576	22,512	57,088	0.42	1.37	0.58
47	Activist	`msg_activist`	101,751	27,501	129,252	1.23	1.67	1.30
48	Public official	`msg_official`	1,372,377	253,098	1,625,475	16.60	15.40	16.40
Events								
49	**Presence of Events (Primary Category)**	`event`	1,531,574	312,091	1,843,665	18.53	18.99	18.61
50	Extreme meteorological event	`evt_weather`	230,498	49,519	280,017	2.79	3.01	2.83
51	Meeting	`evt_meeting`	418,662	100,715	519,377	5.07	6.13	5.24
52	Publication	`evt_publication`	248,794	81,486	330,280	3.01	4.96	3.33
53	Election	`evt_election`	67,362	30,735	98,097	0.81	1.87	0.99
54	New policy	`evt_policy`	409,064	47,286	456,350	4.95	2.88	4.61
55	Judiciary decision	`evt_judiciary`	38,999	7,415	46,414	0.47	0.45	0.47
56	Cultural or Sports event	`evt_cultural`	1,026	25,484	26,510	0.01	1.55	0.27
57	Protest	`evt_protest`	8,078	5,881	13,959	0.10	0.36	0.14
Solutions								
58	**Presence of Solutions (Primary Category)**	`solution`	1,938,846	347,250	2,286,096	23.46	21.13	23.07
59	Mitigation strategy	`sol_mitigation`	1,331,925	222,359	1,554,284	16.11	13.53	15.69
60	Adaptation strategy	`sol_adaptation`	54,500	22,611	77,111	0.66	1.38	0.78

Continued on next page

Supplementary Table S8 – Continued from previous page

#	Category	Code	Count			Proportion (%)		
			EN	**FR**	**ALL**	**EN**	**FR**	**ALL**
EMOTIONAL TONE								
61	Emotion: positive	`tone_positive`	678,078	209,435	887,513	8.20	12.74	8.96
62	Emotion: negative	`tone_negative`	3,161,855	579,082	3,740,937	38.25	35.24	37.75
63	Emotion: neutral	`tone_neutral`	5,218,406	1,185,185	6,403,591	63.14	72.12	64.63
GEOGRAPHIC FOCUS								
64	Mention of Canada	`canada`	3,644,510	656,257	4,300,767	44.09	39.93	43.40
URGENCY TO ACT								
65	Urgency to act	`urgency`	170,813	39,472	210,285	2.07	2.40	2.12
Total Sentences			**8,265,398**	**1,643,378**	**9,908,776**	**100.00**	**100.00**	**100.00**

Note. Counts and proportions are computed live from `CCF_Database.CCF_processed_data` at table-build time; the "ALL" columns aggregate English and French. Proportions are expressed in percent of the total number of sentences in each language. The total sentence counts at the bottom of the table match the data-dictionary entry for `CCF_processed_data` in Supplementary Table S12.

Supplementary Table S9. Per-category inter-coder reliability metrics on the 1,000-sentence gold standard, with the blind phase (sentences 601–1000) reported separately.

#	Category	Code	Cohen's κ (full)	Blind phase (sentences 601–1000)				n
				κ	Krippendorff α	Gwet AC1	% agree	
MAIN FRAMES								
Economic Frame								
1	**Economic Frame (Primary Category)**	`economic_frame`	0.605	0.625	0.625	0.735	0.868	400
2	Negative impacts of climate change on the economy	`eco_neg_impact`	0.694	0.542	0.542	0.915	0.958	400
3	Positive impacts of climate change on the economy	`eco_pos_impact`	0.928	0.831	0.831	0.990	0.995	400
4	Economic disadvantages of climate action	`eco_cost`	0.501	0.471	0.472	0.890	0.945	400
5	Economic benefits of climate action	`eco_benefit`	0.462	0.436	0.433	0.905	0.953	400
6	Carbon footprint of the economic sector	`eco_footprint`	0.354	0.441	0.437	0.820	0.910	400
Health Frame								
7	**Health Frame (Primary Category)**	`health_frame`	0.797	0.865	0.865	0.975	0.988	400
8	Negative impacts of climate change on health	`health_neg_impact`	0.790	0.851	0.851	0.975	0.988	400
9	Health co-benefits of climate action	`health_cobenefit`	0.304	-0.004	-0.004	0.980	0.990	400
Security Frame								
10	**Security Frame (Primary Category)**	`security_frame`	0.777	0.851	0.851	0.975	0.988	400
11	Presence of climate refugees	`security_refugees`	0.826	0.907	0.907	0.990	0.995	400

Continued on next page

Supplementary Table S9 – Continued from previous page

#	Category	Code	Cohen's κ (full)	Blind phase (sentences 601–1000)				n
				κ	Krippendorff α	Gwet AC1	% agree	
12	Conflict	security_conflict	0.844	1.000	1.000	1.000	1.000	400
13	Post-disaster military assistance	security_military	0.283	-0.003	-0.001	0.990	0.995	400
14	Disruption of military operations	security_disruption	0.000	–	1.000	1.000	1.000	400
Justice Frame								
15	**Justice Frame (Primary Category)**	justice_frame	0.638	0.720	0.719	0.890	0.945	400
16	Winners and losers of climate action	justice_winners	0.275	0.386	0.381	0.925	0.963	400
17	Differentiated responsibility	justice_responsibility	0.637	0.655	0.654	0.950	0.975	400
18	Unequal vulnerability to climate change	justice_vulnerability	0.676	0.672	0.672	0.945	0.973	400
19	Unequal access to climate action	justice_access	0.058	0.168	0.160	0.910	0.955	400
20	Intergenerational justice	justice_intergen	0.645	0.731	0.731	0.975	0.988	400
Political Frame								
21	**Political Frame (Primary Category)**	political_frame	0.632	0.664	0.662	0.700	0.850	400
22	Policy action	pol_action	0.445	0.341	0.340	0.825	0.912	400
23	Political debate	pol_debate	0.498	0.542	0.534	0.615	0.807	400
24	Political positioning	pol_position	0.248	0.191	0.187	0.890	0.945	400
25	Public opinion data	pol_opinion	0.709	0.855	0.855	0.990	0.995	400
Scientific Frame								
26	**Scientific Frame (Primary Category)**	scientific_frame	0.579	0.564	0.562	0.770	0.885	400
27	Scientific debate	sci_debate	0.318	0.402	0.393	0.890	0.945	400

Continued on next page

Supplementary Table S9 – Continued from previous page

#	Category	Code	Cohen's κ (full)	Blind phase (sentences 601–1000)				n
				κ	Krippendorff α	Gwet AC1	% agree	
28	Popularisation or scientific discovery	`sci_discovery`	0.506	0.518	0.517	0.805	0.902	400
29	Questioning of climate science	`sci_skepticism`	0.290	0.196	0.191	0.960	0.980	400
30	Defense of climate science	`sci_defense`	0.661	0.748	0.748	0.990	0.995	400
Environmental Frame								
31	**Environmental Frame (Primary Category)**	`environmental_frame`	0.505	0.588	0.587	0.905	0.953	400
32	Loss of natural environments	`env_habitat`	0.358	0.450	0.448	0.910	0.955	400
33	Loss of fauna and flora	`env_species`	0.410	0.432	0.430	0.915	0.958	400
Cultural Frame								
34	**Cultural Frame (Primary Category)**	`cultural_frame`	0.594	0.757	0.758	0.955	0.978	400
35	Artistic representation	`cult_art`	0.696	0.792	0.792	0.970	0.985	400
36	Difficulty to host cultural or sports events	`cult_event_impact`	0.000	0.000	0.993	0.985	0.993	400
37	Loss of indigenous practices	`cult_indigenous`	0.665	1.000	1.000	1.000	1.000	400
38	Carbon footprint of the cultural and sports sectors	`cult_footprint`	0.666	–	1.000	1.000	1.000	400
PRIMARY CATEGORIES								
Actors/Messengers								
39	**Presence of Messengers (Primary Category)**	`messenger`	0.712	0.801	0.801	0.810	0.905	400
40	Health expert	`msg_health`	0.802	0.762	0.762	0.970	0.985	400
41	Economic expert	`msg_economic`	0.680	0.652	0.653	0.905	0.953	400
42	Security expert	`msg_security`	0.490	0.460	0.460	0.955	0.978	400
43	Legal expert	`msg_legal`	0.686	0.886	0.886	0.990	0.995	400
44	Cultural or Sport expert	`msg_cultural`	0.666	0.699	0.699	0.940	0.970	400
45	Natural scientist	`msg_scientist`	0.607	0.692	0.691	0.885	0.943	400

Continued on next page

Supplementary Table S9 – Continued from previous page

#	Category	Code	Cohen's κ (full)	Blind phase (sentences 601–1000)				n
				κ	Krippendorff α	Gwet AC1	% agree	
46	Social scientist	`msg_social`	0.568	0.653	0.653	0.945	0.973	400
47	Activist	`msg_activist`	0.621	0.560	0.561	0.895	0.948	400
48	Public official	`msg_official`	0.609	0.636	0.636	0.755	0.877	400
Events								
49	**Presence of Events (Primary Category)**	`event`	0.667	0.626	0.626	0.630	0.815	400
50	Extreme meteorological event	`evt_weather`	0.700	0.718	0.718	0.945	0.973	400
51	Meeting	`evt_meeting`	0.669	0.762	0.762	0.925	0.963	400
52	Publication	`evt_publication`	0.778	0.807	0.807	0.925	0.963	400
53	Election	`evt_election`	0.732	0.678	0.678	0.955	0.978	400
54	New policy	`evt_policy`	0.502	0.398	0.397	0.785	0.892	400
55	Judiciary decision	`evt_judiciary`	0.810	0.828	0.828	0.980	0.990	400
56	Cultural or Sports event	`evt_cultural`	0.737	0.632	0.631	0.945	0.973	400
57	Protest	`evt_protest`	0.749	0.805	0.805	0.970	0.985	400
Solutions								
58	**Presence of Solutions (Primary Category)**	`solution`	0.699	0.650	0.649	0.785	0.892	400
59	Mitigation strategy	`sol_mitigation`	0.712	0.661	0.661	0.820	0.910	400
60	Adaptation strategy	`sol_adaptation`	0.636	0.556	0.556	0.940	0.970	400
EMOTIONAL TONE								
61	Emotion: positive	`tone_positive`	0.454	0.468	0.468	0.830	0.915	400
62	Emotion: negative	`tone_negative`	0.543	0.561	0.561	0.590	0.795	400
63	Emotion: neutral	`tone_neutral`	0.441	0.468	0.468	0.470	0.735	400
GEOGRAPHIC FOCUS								
64	Mention of Canada	`canada`	0.876	0.870	0.870	0.870	0.935	400
URGENCY TO ACT								
65	Urgency to act	`urgency`	0.176	0.083	0.061	0.895	0.948	400

Note. Cohen's κ, Krippendorff's α, Gwet's AC1, and percent agreement are computed between the expert annotator (gold standard) and the independent second annotator. Only Cohen's κ is reported in both forms—full-sample (all 1,000 doubly-coded sentences) and blind-phase; the other indices are reported for the blind phase only (sentences 601–1000, coded without any communication with the expert), which is the primary reliability benchmark used throughout the paper. Categories with "–" in the blind-phase columns had no positive examples in sentences 601–1000.

Supplementary Table S10. Reliability tier assignment for each of the 65 annotation categories.

#	Category	Code	Validation F_1 macro			Blind	Tier			Flag
			All	EN	FR	κ	All	EN	FR	
MAIN FRAMES										
Economic Frame										
1	**Economic Frame (Primary Category)**	`economic_frame`	0.830	0.847	0.814	0.625	**A**	A	A	
2	Negative impacts of climate change on the economy	`eco_neg_impact`	0.783	0.820	0.741	0.542	**B**	B	B	
3	Positive impacts of climate change on the economy	`eco_pos_impact`	0.684	0.806	0.508	0.831	**B**	A	C	
4	Economic disadvantages of climate action	`eco_cost`	0.723	0.741	0.702	0.471	**B**	B	B	
5	Economic benefits of climate action	`eco_benefit`	0.848	0.885	0.817	0.436	**B**	B	B	
6	Carbon footprint of the economic sector	`eco_footprint`	0.787	0.785	0.788	0.441	**B**	B	B	
Health Frame										
7	**Health Frame (Primary Category)**	`health_frame`	0.808	0.800	0.817	0.865	**A**	A	A	
8	Negative impacts of climate change on health	`health_neg_impact`	0.364	0.386	0.340	0.851	**C**	C	C	
9	Health co-benefits of climate action	`health_cobenefit`	0.186	0.252	0.103	-0.004	**C**	C	C	
Security Frame										
10	**Security Frame (Primary Category)**	`security_frame`	0.782	0.825	0.730	0.851	**B**	A	B	
11	Presence of climate refugees	`security_refugees`	0.481	0.476	0.417	0.907	**C**	C	C	

Continued on next page

Supplementary Table S10 – Continued from previous page

#	Category	Code	Validation F_1 macro			Blind	Tier			Flag
			All	EN	FR	κ	All	EN	FR	
12	Conflict	security_conflict	0.324	0.358	0.286	1.000	**C**	C	C	
13	Post-disaster military assistance	security_military	0.045	0.000	0.045	-0.003	**C**	C	C	no EN model
14	Disruption of military operations	security_disruption	1.000	0.000	1.000	–	**C**	C	C	no EN model
Justice Frame										
15	**Justice Frame (Primary Category)**	justice_frame	0.857	0.832	0.880	0.720	**A**	A	A	
16	Winners and losers of climate action	justice_winners	0.708	0.654	0.737	0.386	**C**	C	C	PIRD*
17	Differentiated responsibility	justice_responsibility	0.860	0.900	0.823	0.655	**A**	A	A	
18	Unequal vulnerability to climate change	justice_vulnerability	0.845	0.786	0.894	0.672	**A**	B	A	
19	Unequal access to climate action	justice_access	0.651	0.475	0.796	0.168	**C**	C	C	
20	Intergenerational justice	justice_intergen	0.942	0.924	0.952	0.731	**A**	A	A	
Political Frame										
21	**Political Frame (Primary Category)**	political_frame	0.829	0.812	0.842	0.664	**A**	A	A	
22	Policy action	pol_action	0.756	0.802	0.722	0.341	**C**	C	C	PIRD*
23	Political debate	pol_debate	0.564	0.469	0.650	0.542	**C**	C	B	
24	Political positioning	pol_position	0.782	0.752	0.808	0.191	**C**	C	C	PIRD*
25	Public opinion data	pol_opinion	0.894	0.877	0.910	0.855	**A**	A	A	
Scientific Frame										
26	**Scientific Frame (Primary Category)**	scientific_frame	0.890	0.891	0.890	0.564	**B**	B	B	

Continued on next page

Supplementary Table S10 – Continued from previous page

#	Category	Code	Validation F_1 macro			Blind	Tier			Flag
			All	EN	FR	κ	All	EN	FR	
27	Scientific debate	sci_debate	0.911	0.962	0.867	0.402	**B**	B	B	
28	Popularisation or scientific discovery	sci_discovery	0.838	0.884	0.797	0.518	**B**	B	B	
29	Questioning of climate science	sci_skepticism	0.775	0.783	0.768	0.196	**C**	C	C	PIRD*
30	Defense of climate science	sci_defense	0.682	0.769	0.460	0.748	**B**	B	C	
Environmental Frame										
31	**Environmental Frame (Primary Category)**	environmental_frame	0.871	0.866	0.875	0.588	**B**	B	B	
32	Loss of natural environments	env_habitat	0.721	0.804	0.573	0.450	**B**	B	C	
33	Loss of fauna and flora	env_species	0.703	0.714	0.692	0.432	**B**	B	B	
Cultural Frame										
34	**Cultural Frame (Primary Category)**	cultural_frame	0.708	0.802	0.581	0.757	**B**	A	C	
35	Artistic representation	cult_art	0.659	0.834	0.469	0.792	**B**	A	C	
36	Difficulty to host cultural or sports events	cult_event_impact	0.484	0.631	0.354	0.000	**C**	C	C	
37	Loss of indigenous practices	cult_indigenous	0.595	0.824	0.415	1.000	**C**	A	C	
38	Carbon footprint of the cultural and sports sectors	cult_footprint	0.351	0.702	0.014	–	**C**	B	C	
PRIMARY CATEGORIES										
Actors/Messengers										
39	**Presence of Messengers (Primary Category)**	messenger	0.961	0.955	0.967	0.801	**A**	A	A	
40	Health expert	msg_health	0.907	0.862	0.951	0.762	**A**	A	A	
41	Economic expert	msg_economic	0.874	0.847	0.895	0.652	**A**	A	A	

Continued on next page

Supplementary Table S10 – Continued from previous page

#	Category	Code	Validation F_1 macro			Blind	Tier			Flag
			All	EN	FR	κ	All	EN	FR	
42	Security expert	`msg_security`	0.787	0.728	0.813	0.460	**B**	B	B	
43	Legal expert	`msg_legal`	0.914	0.861	0.927	0.886	**A**	A	A	
44	Cultural or Sport expert	`msg_cultural`	0.843	0.853	0.825	0.699	**A**	A	A	
45	Natural scientist	`msg_scientist`	0.907	0.920	0.896	0.692	**A**	A	A	
46	Social scientist	`msg_social`	0.944	0.938	0.948	0.653	**A**	A	A	
47	Activist	`msg_activist`	0.918	0.934	0.903	0.560	**B**	B	B	
48	Public official	`msg_official`	0.929	0.927	0.930	0.636	**A**	A	A	
Events										
49	**Presence of Events (Primary Category)**	`event`	0.897	0.881	0.912	0.626	**A**	A	A	
50	Extreme meteorological event	`evt_weather`	0.889	0.887	0.891	0.718	**A**	A	A	
51	Meeting	`evt_meeting`	0.877	0.855	0.891	0.762	**A**	A	A	
52	Publication	`evt_publication`	0.958	0.950	0.962	0.807	**A**	A	A	
53	Election	`evt_election`	0.903	0.911	0.895	0.678	**A**	A	A	
54	New policy	`evt_policy`	0.829	0.821	0.837	0.398	**C**	C	C	PIRD*
55	Judiciary decision	`evt_judiciary`	0.808	0.628	0.939	0.828	**A**	B	A	
56	Cultural or Sports event	`evt_cultural`	0.881	0.906	0.857	0.632	**A**	A	A	
57	Protest	`evt_protest`	0.943	0.982	0.909	0.805	**A**	A	A	
Solutions										
58	**Presence of Solutions (Primary Category)**	`solution`	0.900	0.891	0.909	0.650	**A**	A	A	
59	Mitigation strategy	`sol_mitigation`	0.839	0.872	0.801	0.661	**A**	A	A	
60	Adaptation strategy	`sol_adaptation`	0.895	0.889	0.903	0.556	**B**	B	B	
EMOTIONAL TONE										
61	Emotion: positive	`tone_positive`	0.702	0.712	0.692	0.468	**B**	B	B	
62	Emotion: negative	`tone_negative`	0.799	0.793	0.804	0.561	**B**	B	B	
63	Emotion: neutral	`tone_neutral`	0.777	0.756	0.798	0.468	**B**	B	B	

Continued on next page

Supplementary Table S10 – Continued from previous page

#	Category	Code	Validation F_1 macro			Blind	Tier			Flag
			All	EN	FR	κ	All	EN	FR	
GEOGRAPHIC FOCUS										
64	Mention of Canada	`canada`	0.982	0.986	0.978	0.870	**A**	A	A	
URGENCY TO ACT										
65	Urgency to act	`urgency`	0.853	0.874	0.835	0.083	**C**	C	C	PIRD*
TIER COUNTS: All: A=27 / B=21 / C=17; EN: A=30 / B=20 / C=15; FR: A=27 / B=17 / C=21; PIRD*: 6; without an English model: 2.										

Note. Each category is assigned to a reliability tier based on its validation F_1 macro and the blind-phase Cohen's κ. **Tier A** (research-grade) requires $F_1 \geq 0.80$ *and* $\kappa \geq 0.60$; **Tier C** (exploratory) corresponds to $F_1 < 0.60$ *or* $\kappa < 0.40$; **Tier B** (use with caution) captures everything in between. The acronym **PIRD*** in the Flag column stands for *prevalence-induced reliability deflation*: it identifies categories with $F_1 \geq 0.70$ but blind-phase Cohen's $\kappa < 0.40$, where the chance-corrected agreement is depressed by very low prevalence rather than by systematic disagreement between coders [19, 20]. The two categories without an English model (`security_military` and `security_disruption`, annotated on French sentences only for lack of positive English training examples) are flagged "no EN model" and forced to Tier C in every language column; the `exclusion_reason` column of the deposited lookup carries the value `insufficient_training_data` for these two rows and is `NULL` everywhere else.

Supplementary Table S11. Per-model training provenance: best epoch, training phase, validation macro F_1, and positive-class support for every BERT/CamemBERT classifier.

#	Category	Code	English (BERT)				French (CamemBERT)			
			Epoch	Training phase	F_1	n^+	Epoch	Training phase	F_1	n^+
MAIN FRAMES										
Economic Frame										
1	**Economic Frame (Primary Category)**	`economic_frame`	4	r	0.845	24	13	n	0.885	28
2	Negative impacts of climate change on the economy	`eco_neg_impact`	12	n	0.821	6	3	n	0.917	8
3	Positive impacts of climate change on the economy	`eco_pos_impact`	6	n	0.664	5	4	r	0.698	3
4	Economic disadvantages of climate action	`eco_cost`	5	n	0.732	6	1	r	0.707	4
5	Economic benefits of climate action	`eco_benefit`	9	n	0.726	3	17	n	0.749	16
6	Carbon footprint of the economic sector	`eco_footprint`	8	n	0.897	7	16	n	0.904	8
Health Frame										
7	**Health Frame (Primary Category)**	`health_frame`	9	n	0.894	6	6	n	0.830	4
8	Negative impacts of climate change on health	`health_neg_impact`	1	n	0.455	5	1	n	0.429	3
9	Health co-benefits of climate action	`health_cobenefit`	17	r	0.486	1	2	r	0.119	2
Security Frame										
10	**Security Frame (Primary Category)**	`security_frame`	9	n	0.933	12	11	n	0.898	2
11	Presence of climate refugees	`security_refugees`	2	n	1.000	1	9	r	1.000	1
12	Conflict	`security_conflict`	1	n	1.000	1	14	r	1.000	1
13	Post-disaster military assistance	`security_military`	–	–	–	–	3	r	0.026	2
14	Disruption of military operations	`security_disruption`	–	–	–	–	18	n	1.000	1
Justice Frame										
15	**Justice Frame (Primary Category)**	`justice_frame`	5	n	0.847	32	10	n	0.849	30

Continued on next page

Supplementary Table S11 – Continued from previous page

#	Category	Code	English (BERT)				French (CamemBERT)			
			Epoch	Training phase	F_1	n^+	Epoch	Training phase	F_1	n^+
16	Winners and losers of climate action	`justice_winners`	13	n	0.800	2	14	n	0.795	2
17	Differentiated responsibility	`justice_responsibility`	4	n	0.883	4	3	n	0.750	4
18	Unequal vulnerability to climate change	`justice_vulnerability`	19	r	0.782	6	13	n	0.798	5
19	Unequal access to climate action	`justice_access`	13	r	0.182	2	4	r	0.809	9
20	Intergenerational justice	`justice_intergen`	9	n	0.966	7	11	n	0.899	3
Political Frame										
21	**Political Frame (Primary Category)**	`political_frame`	3	n	0.853	46	5	n	0.831	133
22	Policy action	`pol_action`	2	n	0.803	17	4	n	0.816	6
23	Political debate	`pol_debate`	7	n	0.569	38	4	n	0.768	43
24	Political positioning	`pol_position`	5	n	0.848	6	4	n	0.895	3
25	Public opinion data	`pol_opinion`	4	n	0.954	6	12	n	1.000	2
Scientific Frame										
26	**Scientific Frame (Primary Category)**	`scientific_frame`	3	n	0.869	27	5	n	0.832	56
27	Scientific debate	`sci_debate`	5	n	0.793	10	7	n	0.688	5
28	Popularisation or scientific discovery	`sci_discovery`	3	n	0.814	14	6	n	0.844	16
29	Questioning of climate science	`sci_skepticism`	3	r	0.827	5	5	r	0.664	4
30	Defense of climate science	`sci_defense`	6	n	0.752	4	1	n	0.459	3
Environmental Frame										
31	**Environmental Frame (Primary Category)**	`environmental_frame`	7	n	0.915	9	13	r	0.802	6
32	Loss of natural environments	`env_habitat`	7	n	1.000	5	5	n	0.829	3
33	Loss of fauna and flora	`env_species`	5	n	0.873	4	3	n	0.829	3
Cultural Frame										
34	**Cultural Frame (Primary Category)**	`cultural_frame`	7	n	0.879	24	13	r	0.913	5
35	Artistic representation	`cult_art`	2	n	1.000	2	7	n	1.000	2
36	Difficulty to host cultural or sports events	`cult_event_impact`	18	n	0.850	7	4	n	1.000	2

Continued on next page

Supplementary Table S11 – Continued from previous page

#	Category	Code	English (BERT)				French (CamemBERT)			
			Epoch	**Training phase**	F_1	n^+	**Epoch**	**Training phase**	F_1	n^+
37	Loss of indigenous practices	`cult_indigenous`	6	n	1.000	1	3	r	0.462	2
38	Carbon footprint of the cultural and sports sectors	`cult_footprint`	7	n	1.000	1	3	r	0.074	1
PRIMARY CATEGORIES										
Actors/Messengers										
39	**Presence of Messengers (Primary Category)**	`messenger`	11	n	0.908	72	6	n	0.922	227
40	Health expert	`msg_health`	9	r	0.927	6	8	n	0.954	6
41	Economic expert	`msg_economic`	4	n	0.785	6	3	r	0.861	7
42	Security expert	`msg_security`	8	n	0.830	2	9	n	1.000	1
43	Legal expert	`msg_legal`	4	n	0.833	2	11	n	0.883	6
44	Cultural or Sport expert	`msg_cultural`	3	n	1.000	1	3	r	0.773	9
45	Natural scientist	`msg_scientist`	8	n	0.883	37	8	n	0.902	12
46	Social scientist	`msg_social`	7	n	0.883	7	16	r	0.816	12
47	Activist	`msg_activist`	5	n	0.775	5	5	n	1.000	6
48	Public official	`msg_official`	4	n	0.800	18	10	n	0.943	24
Events										
49	**Presence of Events (Primary Category)**	`event`	11	n	0.865	33	10	n	0.876	113
50	Extreme meteorological event	`evt_weather`	3	n	1.000	8	7	n	0.912	6
51	Meeting	`evt_meeting`	8	n	0.880	8	3	n	0.969	11
52	Publication	`evt_publication`	4	n	0.931	25	7	n	1.000	12
53	Election	`evt_election`	4	n	1.000	1	10	r	0.899	5
54	New policy	`evt_policy`	7	n	0.856	6	8	n	0.841	6
55	Judiciary decision	`evt_judiciary`	7	n	1.000	1	9	n	1.000	1
56	Cultural or Sports event	`evt_cultural`	4	n	0.664	5	7	n	1.000	1
57	Protest	`evt_protest`	4	n	0.944	5	5	n	0.862	6
Solutions										
58	**Presence of Solutions (Primary Category)**	`solution`	6	n	0.825	90	4	n	0.911	46
59	Mitigation strategy	`sol_mitigation`	16	n	0.842	79	7	n	0.877	105
60	Adaptation strategy	`sol_adaptation`	9	n	0.843	15	10	n	0.894	21
EMOTIONAL TONE										
61	Emotion: positive	`tone_positive`	8	n	0.746	11	6	n	0.829	16
62	Emotion: negative	`tone_negative`	2	n	0.745	152	16	n	0.804	158

Continued on next page

Supplementary Table S11 – Continued from previous page

#	Category	Code	English (BERT)				French (CamemBERT)			
			Epoch	Training phase	F_1	n^+	Epoch	Training phase	F_1	n^+
63	Emotion: neutral	`tone_neutral`	4	n	0.709	205	4	n	0.741	80
GEOGRAPHIC FOCUS										
64	Mention of Canada	`canada`	7	n	0.955	136	12	n	0.972	54
URGENCY TO ACT										
65	Urgency to act	`urgency`	10	r	0.783	20	7	n	0.816	19

Summary: 128 trained models. Reinforcement triggered for 45 models; best epoch retained from the reinforced phase for 23 models. Static training configuration (optimiser, learning rate, batch size, random seed, hardware) is reported in the Methods section.

Note. The English models are BERT-base-cased (case-sensitive: load their tokeniser with `do_lower_case=False`) and the French models are CamemBERT-base, both trained through the AugmentedSocialScientist framework. The four per-language sub-columns are: **Ep.** (best epoch retained by the combined-criterion best-model selector $0.98 \cdot F_1^{(1)} + 0.02 \cdot$ macro-F_1), **Ph.** (training phase — n for normal, r for reinforced training, the latter firing when the positive-class F_1 falls below 0.70 during normal training), F_1 (macro F_1 at the retained epoch), and n^+ (number of positive validation examples; "–" denotes models that failed to train because of insufficient positive examples). The static configuration (optimiser AdamW with $\epsilon = 10^{-8}$, learning rate 5×10^{-5} in normal training and 5×10^{-6} in the reinforced phase, batch size 32 normal / 64 reinforced, maximum sequence length 512 tokens, up to 20 epochs in each phase, random seed 42 for Python/NumPy/PyTorch RNGs, linear schedule with zero warm-up, an unweighted cross-entropy in normal training and, in the reinforced phase, a `WeightedRandomSampler` combined with a class-weighted cross-entropy (positive-class weight 3.5), training on an Apple Mac Studio M2 Ultra with 128 GB unified memory) is documented in full in the Methods section of the main manuscript.

Supplementary Table S12. Complete data dictionary of the enriched CCF Database: every column of every table, its type, and a one-line definition.

Column	Type	Definition
Table CCF_full_data *(article-level metadata, 283,964 rows)*		
`doc_id`	BIGINT (PK)	Unique article identifier.
`news_type`	TEXT	Editorial category (news, opinion, editorial, etc.).
`title`	TEXT	Article headline.
`author`	TEXT	Author byline when available.
`media`	TEXT	Newspaper name.
`words_count`	BIGINT	Total word count of the article.
`date`	DATE	Publication date.
`language`	TEXT	"EN" or "FR".
`page_number`	TEXT	Print page (free text, e.g. "A1" or "B3").
Table CCF_processed_data *(sentence-level annotations, 9,908,776 rows)*		
`doc_id`	BIGINT	Foreign key to CCF_full_data.
`sentence_id`	BIGINT	Sentence position within the article (starting at 1; 0 reserved for title in the embeddings table).
Ten article-level metadata columns are duplicated here for query convenience (news_type, title, author, media, words_count, date, language, page_number).		
65 annotation columns	INTEGER (0/1, nullable)	One binary flag per category from Supplementary Table S3. Sub-categories are NULL when their parent primary detection is negative.
`ner_entities`	TEXT (JSON)	Per-sentence NER output as `{"PER": [..], "ORG": [..], "LOC": [..]}`.
Table CCF_article_aggregates *(article-level rollup, 283,964 rows)*		
`doc_id`	BIGINT (PK)	Foreign key to CCF_full_data.
`n_sentences`	INTEGER	Total number of two-sentence analytical units in the article.
`n_sentences_en`	INTEGER	Subset in English.
`n_sentences_fr`	INTEGER	Subset in French.
65 `prop_X` columns	DOUBLE PRECISION ($[0, 1]$)	Article-level proportion of sentences positive for category X (sum / n_sentences, with NULL sub-categories coalesced to 0).
`top_frame`	TEXT	Main frame with the highest `prop_X`, or NULL for articles with no positive frame.
`top_frame_prop`	DOUBLE PRECISION	Proportion of the top frame.
`entropy_frames`	DOUBLE PRECISION	Shannon entropy (natural base) of the eight main-frame proportions; 0 when the article has no positive frame.

Continued on next page

Supplementary Table S12 – Continued from previous page

Column	Type	Definition
n_unique_messenger_subs	SMALLINT	Number of distinct messenger sub-categories present in the article (0–9).
n_unique_event_subs	SMALLINT	Same for event sub-categories (0–8).
n_unique_solution_subs	SMALLINT	Same for solution sub-categories (0–2).
Table CCF_article_entities *(NER rollup, 283,964 rows)*		
doc_id	BIGINT (PK)	Foreign key to CCF_full_data.
entities_per	JSONB	Deduplicated array of persons mentioned in the article.
entities_org	JSONB	Deduplicated array of organisations.
entities_loc	JSONB	Deduplicated array of locations.
n_unique_per	INTEGER	Cardinality of entities_per.
n_unique_org	INTEGER	Cardinality of entities_org.
n_unique_loc	INTEGER	Cardinality of entities_loc.
n_sentences_with_per	INTEGER	Number of sentences with at least one PER entity.
n_sentences_with_org	INTEGER	Same for ORG.
n_sentences_with_loc	INTEGER	Same for LOC.
Table CCF_reliability_tiers *(per-category quality lookup, 65 rows)*		
number	SMALLINT (PK)	Table B1 number (1–65).
category	TEXT	Readable category name.
code	TEXT (UNIQUE)	Column code as used in CCF_processed_data.
tier_overall	CHAR(1)	Tier on the combined corpus: “A”, “B”, or “C”.
tier_en	CHAR(1)	Tier on the English subset.
tier_fr	CHAR(1)	Tier on the French subset.
paradox_kappa	BOOLEAN	True if the category exhibits prevalence-induced reliability deflation, i.e. $F_1 \geq 0.70$ and $\kappa < 0.40$. The column name is kept for backwards compatibility with the original schema.
exclusion_reason	TEXT	“insufficient_training_data” for the two categories that could not be trained, NULL otherwise.

Continued on next page

Supplementary Table S12 – Continued from previous page

Column	Type	Definition
`f1_macro_all`, `f1_macro_en`, `f1_macro_fr`	DOUBLE PRECISION	Validation macro F_1 score on the gold standard.
`support_all`, `support_en`, `support_fr`	INTEGER	Number of positive (class 1) validation examples.
`cohens_kappa_blind_phase`, `gwet_ac1_blind_phase`, `krippendorff_alpha_blind_phase`	DOUBLE PRECISION	Inter-coder agreement on the blind phase (sentences 601–1000).
`n_blind_phase`	INTEGER	Sentence count of the blind phase.
`prev_coder1_full`, `prev_coder2_full`	DOUBLE PRECISION	Prevalence (proportion of positives) per coder on the full 1,000-sentence sample.
Table CCF_sentence_embeddings *(BGE-M3 dense vectors, 10,192,740 rows)*		
`doc_id`	BIGINT (PK, with `sentence_id`)	Foreign key to CCF_full_data.
`sentence_id`	INTEGER (PK)	Sentence position within the article. `sentence_id` = 0 is reserved for the article title.
`embedding`	`halfvec(1024)`	L2-normalised BGE-M3 vector. `pgvector` cosine distance via `<=>`; HNSW index on cosine ops.

Note. All annotation column codes (`economic_frame`, `eco_neg_impact`, etc.) and their definitions appear in Supplementary Table S3.

Supplementary references

[1] Jennifer Good. "The Framing of Climate Change in Canadian, American, and International Newspapers: A Media Propaganda Model Analysis". In: *Canadian Journal of Communication* 33.2 (June 2008). Publisher: University of Toronto Press, pp. 233–256. ISSN: 0705-3657. DOI: 10.22230/cjc.2008v33n2a2017. URL: https://cjc.utppublishing.com/doi/full/10.22230/cjc.2008v33n2a2017.

[2] Dan Rowe. "Comparing Newspaper Coverage of Climate Change During Election Campaigns in the United States, Canada and Australia". Anglais. ISBN: 9781267179050. Ph.D. 2011. URL: https://www.proquest.com/docview/922669062/abstract/2CFD056A38334D6EPQ/1.

[3] Nathan Young and Eric Dugas. "Representations of Climate Change in Canadian National Print Media: The Banalization of Global Warming". en. In: *Canadian Review of Sociology/Revue canadienne de sociologie* 48.1 (Feb. 2011). Publisher: John Wiley & Sons, Ltd, pp. 1–22. ISSN: 1755-618X. DOI: 10.1111/j.1755-618X.2011.01247.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1755-618X.2011.01247.x.

[4] Darryn Anne DiFrancesco and Nathan Young. "Seeing climate change: the visual construction of global warming in Canadian national print media". EN. In: *cultural geographies* 18.4 (Oct. 2011). Publisher: SAGE Publications Ltd, pp. 517–536. ISSN: 1474-4740. DOI: 10.1177/1474474010382072. URL: https://doi.org/10.1177/1474474010382072.

[5] Katrina Ahchong and Rachel Dodds. "Anthropogenic climate change coverage in two Canadian newspapers, the *Toronto Star* and the *Globe and Mail*, from 1988 to 2007". In: *Environmental Science & Policy* 15.1 (Jan. 2012), pp. 48–59. ISSN: 1462-9011. DOI: 10.1016/j.envsci.2011.09.008. URL: https://www.sciencedirect.com/science/article/pii/S146290111100150X.

[6] Nathan Young and Eric Dugas. "Comparing Climate Change Coverage in Canadian English- and French-Language Print Media: Environmental Values, Media Cultures, and the Narration of Global Warming". In: *The Canadian Journal of Sociology / Cahiers canadiens de sociologie* 37.1 (2012). Publisher: Canadian Journal of Sociology, pp. 25–54. ISSN: 0318-6431. DOI: 10.29173/cjs9733. URL: https://doi.org/10.29173/cjs9733.

[7] Mark Stoddart and David Tindall. "Canadian news media and the cultural dynamics of multilevel climate governance". In: *Environmental Politics* 24.3 (May 2015). Publisher: Routledge _eprint: https://doi.org/10.1080/09644016.2015.1008249, pp. 401–422. ISSN: 0964-4016. DOI: 10.1080/09644016.2015.1008249. URL: https://doi.org/10.1080/09644016.2015.1008249.

[8] James Ford and Diana King. "Coverage and framing of climate change adaptation in the media: A review of influential North American newspapers during 1993–2013". In: *Environmental Science & Policy* 48 (Apr. 2015), pp. 137–146. ISSN: 1462-9011. DOI: 10.1016/j.envsci.2014.12.003. URL: https://www.sciencedirect.com/science/article/pii/S1462901114002317.

[9] Mark Stoddart and Jillian Smith. "The Endangered Arctic, the Arctic as Resource Frontier: Canadian News Media Narratives of Climate Change and the North". en. In: *Canadian Review of Sociology/Revue canadienne de sociologie* 53.3 (Aug. 2016). Publisher: John Wiley & Sons, Ltd, pp. 316–336. ISSN: 1755-618X. DOI: 10.1111/cars.12111. URL: https://onlinelibrary.wiley.com/doi/10.1111/cars.12111.

[10] Mark Stoddart et al. "Media Access and Political Efficacy in the Eco-politics of Climate Change: Canadian National News and Mediated Policy Networks". In: *Environmental Communication* 11.3 (May 2017). Publisher: Routledge _eprint: https://doi.org/10.1080/17524032.2016.1275731, pp. 386–400. ISSN: 1752-4032. DOI: 10.1080/17524032.2016.1275731. URL: https://doi.org/10.1080/17524032.2016.1275731.

[11] Ralf Barkemeyer et al. "Media coverage of climate change: An international comparison". EN. In: *Environment and Planning C: Politics and Space* 35.6 (Sept. 2017). Publisher: SAGE Publications Ltd STM, pp. 1029–1054. ISSN: 2399-6544. DOI: 10.1177/0263774X16680818. URL: https://doi.org/10.1177/0263774X16680818.

[12] Nia King et al. "How do Canadian media report climate change impacts on health? A newspaper review". en. In: *Climatic Change* 152.3 (Mar. 2019), pp. 581–596. ISSN: 1573-1480. DOI: 10.1007/s10584-018-2311-2. URL: https://doi.org/10.1007/s10584-018-2311-2.

[13] Alizee Pillod. "Reframing climate change as a public health issue : a Canadian case study, 2008-2020". fr. PhD thesis. July 2021. URL: http://hdl.handle.net/1866/26126.

[14] Oliver Brandt. "News Media Framing of Climate Change in Canada: A Provincial Analysis of 2015–2018". Anglais. ISBN: 9798496553797. MA thesis. 2021. URL: https://www.proquest.com/docview/2607337799/abstract/16DEF7A383FD4B8DPQ/1.

[15] Valerie Hase et al. "Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018)". In: *Global Environmental Change* 70 (Sept. 2021), p. 102353. ISSN: 0959-3780. DOI: 10.1016/j.gloenvcha.2021.102353. URL: https://www.sciencedirect.com/science/article/pii/S0959378021001321.

[16] Sonya Sachdeva and Sarah McCaffrey. "Themes and patterns in print media coverage of wildfires in the USA, Canada and Australia: 1986–2016". In: *International Journal of Wildland Fire* 31.12 (Oct. 2022), pp. 1089–1102. ISSN: 1049-8001. DOI: 10.1071/WF22174. URL: https://doi.org/10.1071/WF22174.

[17] Mark Stoddart and Yixi Yang. "What are the roles of regional and local climate governance discourse and actors? Mediated climate change policy networks in Atlantic Canada". en. In: *Review of Policy Research* 40.6 (Nov. 2023). Publisher: John Wiley & Sons, Ltd, pp. 1144–1168. ISSN: 1541-1338. DOI: 10.1111/ropr.12510. URL: https://onlinelibrary.wiley.com/doi/10.1111/ropr.12510.

[18] Kaiping Chen et al. “How Climate Movement Actors and News Media Frame Climate Change and Strike: Evidence from Analyzing Twitter and News Media Discourse from 2018 to 2021”. EN. In: *The International Journal of Press/Politics* 28.2 (Apr. 2023). Publisher: SAGE Publications Inc, pp. 384–413. ISSN: 1940-1612. DOI: 10.1177/19401612221106405. URL: https://doi.org/10.1177/19401612221106405.

[19] Alvan R. Feinstein and Domenic V. Cicchetti. “High agreement but low kappa: I. The problems of two paradoxes”. In: *Journal of Clinical Epidemiology* 43.6 (1990), pp. 543–549. DOI: 10.1016/0895-4356(90)90158-L. URL: https://doi.org/10.1016/0895-4356(90)90158-L.

[20] Ted Byrt, Janet Bishop, and John B. Carlin. “Bias, prevalence and kappa”. In: *Journal of Clinical Epidemiology* 46.5 (1993), pp. 423–429. DOI: 10.1016/0895-4356(93)90018-V. URL: https://doi.org/10.1016/0895-4356(93)90018-V.