UNIVERSITY OF LAUSANNE · INSTITUTE OF GEOGRAPHY AND SUSTAINABILITY · TEAM M3 · 29 JUNE 2026

CCF & LLM Tool

Machine learning for climate research and social science

CCF Database · LLM Tool · 20 newspapers · 1978-2024 · EN + FR · open

Alizée Pillod[1], Antoine Lemor[2]

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

CAPP · CIRST · RFICS

ccf-project.ca

LLM Tool

UNIL · M3 LAUSANNE

CAPP Centre d'analyse des politiques publiques

RFICS

CIRST Centre interuniversitaire de recherche sur la science et la technologie

UDS Université de Sherbrooke

Outline

1
Context
climate framing and what existing tools miss

2
CCF Database
corpus · 65 categories · applications

3
LLM Tool
the local hybrid pipeline that produced the labels

4
Conclusion
what travels beyond the climate case

PART I

Context

Two summers, two ways of telling the same crisis

Toronto Star

SUMMER 1988

Summer of '88 could be hint of the future. Scientists fear warming trend will continue. The four hottest years in Toronto's history have all been in the 1980s. By about 2040 or 2050 we'll be well into it.

Globe and Mail

SUMMER 2023

Poilievre seizes the moment on Atlantic Canada carbon-tax fears. (...) The Conservative leader said the wildfires proved that Trudeau's carbon tax was not working. (...) Carbon tax is a loser for Liberals.

Same type of events, but different dominant actors and frames.

RESEARCH QUESTION

If two summer reports can read so differently, what would it take to measure climate discourse, actors and frames across every outlet, every region, every year?

Climate framing matters,
but our measurement tools have not kept up

How issues are framed in the news shapes **what publics believe, what coalitions form, and what policies become thinkable**.

Entman 1993 · Boykoff & Boykoff 2007 · Nisbet 2009

Climate communication research has documented framing dynamics through *« hand coding »* of small samples, or *« topic modelling »* of large but unsupervised corpora.

Stecula & Merkley 2019 · Bohr 2020 · Tvinnereim & Fløttum 2015

Canadian climate discourse is fragmented, bilingual, regional. Existing datasets cover **short windows, single outlets, English only**.

Young & Dugas 2011 · Stoddart et al. 2017 · Boursier & Lemor 2024

No existing corpus combines **multi-decade coverage**, **multi-outlet breadth**, **bilingual EN/FR** and **sentence-level supervised annotation** of frames, actors, events and locations.

What this puzzle demands of a research infrastructure

1 EXTRACT EVERY ACTOR

Persons, organizations, types of messengers (scientists, politicians, activists, experts) named or quoted in every sentence.

2 EXTRACT EVERY FRAME

Economic, political, scientific, environmental, health, justice, cultural, security — each with sub-categories for impacts, debate and tone.

3 ANCHOR EVERY CLAIM

Geographic anchor (provinces, cities), events (disasters, conferences, elections), solutions (mitigation, adaptation) and emotional tone.

Research question

Can we build a multi-decade, multi-outlet, bilingual climate corpus where every sentence has been parsed for actors, frames, events and locations, at near-human accuracy?

PART II

CCF Database

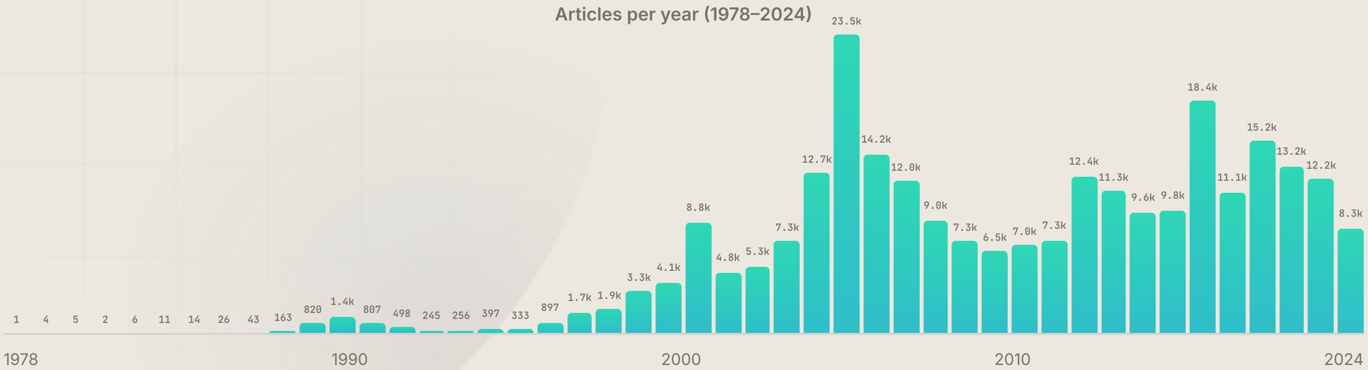
The CCF Database: Canadian Climate Framing

266K
Articles
1978 – 2024

20
Newspapers
EN + FR

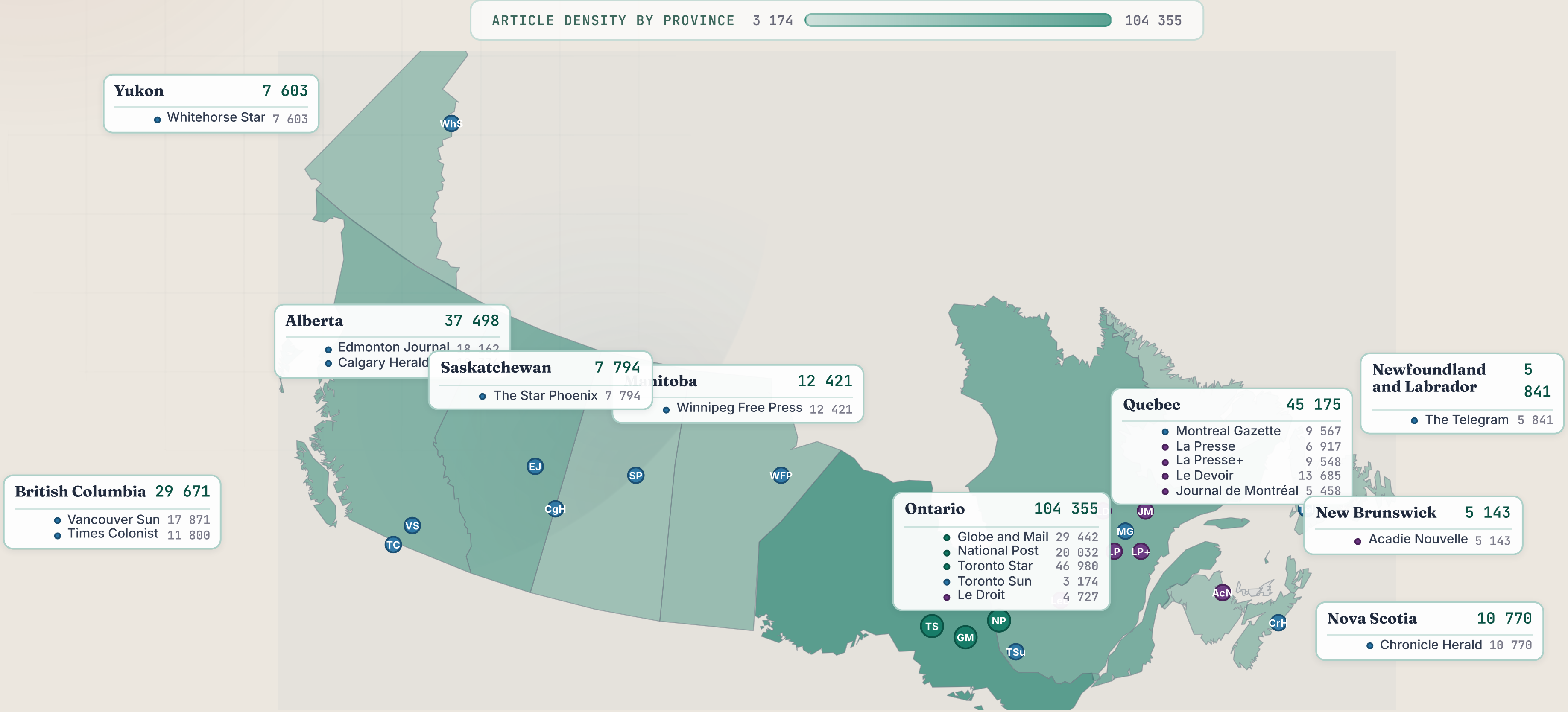
9.2M
Annotated sentences
BERT / CamemBERT

65
Categories
hierarchical



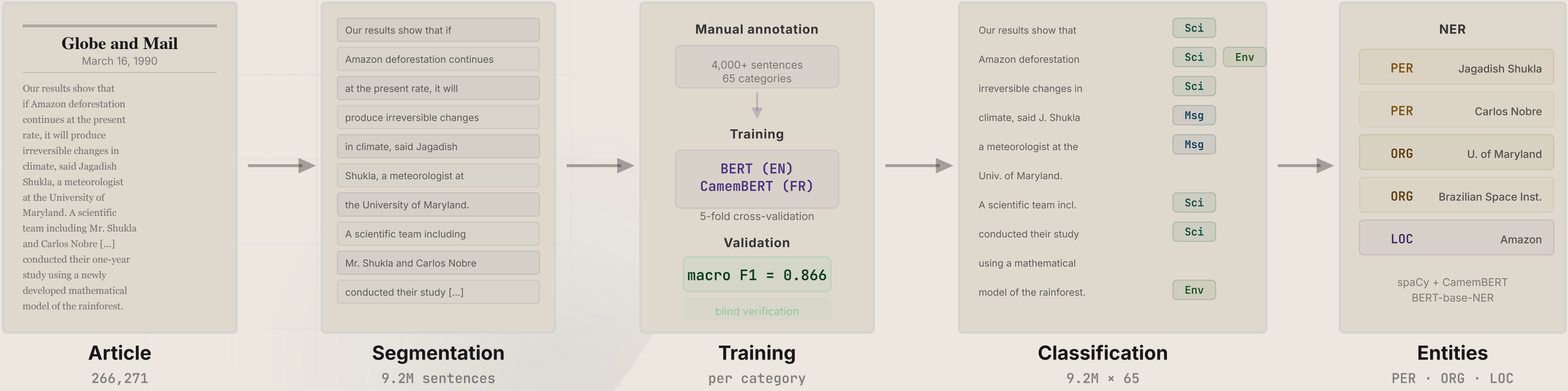
The largest ML-annotated climate discourse corpus in Canada
Macro F1 : 0.866 – Lemor, Pillod, Taylor (2025)

All 20 newspapers in the corpus, mapped



Our method in a nutshell

From a single article to 9.2 million annotated sentences across 65 binary categories



65 CATEGORIES ACROSS 10 FAMILIES

Family	Count	Examples
Economic	6	Costs, benefits, impacts, sector footprint
Political	5	Policy, debate, positioning, opinion
Scientific	5	Discovery, debate, skepticism, defense
Environmental	3	Habitat loss, species loss
Health	3	Negative impacts, co-benefits
Security	5	Refugees, conflict, military, disruption
Justice	6	Winners/losers, responsibility, vulnerability
Cultural	5	Art, events, Indigenous practices, footprint
Messengers	10	Scientists, officials, activists, experts
Events + Solutions	11	Weather, elections, protests, mitigation

+ Emotional tone (3) · Geographic focus (1) · Urgency (1) · NER (PER, ORG, LOC)

65 categories, organised by frame

Annotation in practice

One sentence, four article-level variables: parent frame, two economic sub-categories, and a geographic anchor

Winnipeg Free Press • 21 July 2004 • Manitoba

« By 2050, an average summer temperature in Edmonton could be 1.5 to 4.5 °C higher than present. While Canadians may relish the warmer and longer growing seasons and the lower winter heating costs, the downside will be raging forest fires, insect plagues and heat-stressed livestock. »

LOCATION	FRAME	COSTS	BENEFITS
Edmonton	**Economic frame**	**Costs of climate action**	**Benefits of climate action**
`Alberta`	`eco_frame`	`eco_cost`	`eco_benefit`

One sentence, four article-level variables: a parent frame and its two economic sub-categories, plus a geographic anchor.

`Location` `eco_frame` `eco_cost` `eco_benefit`

Four examples of questions we can answer

1 THEMATIC FRAMES

How have thematic frames evolved across nearly five decades of Canadian climate coverage, since 1978?

2 REGIONAL PATTERNS

How do regional differences in fossil fuel dependence shape media coverage of climate change in Canada?

3 POLICYMAKER NETWORK

Which policymakers structure Canadian climate discourse today, and who appears together with whom?

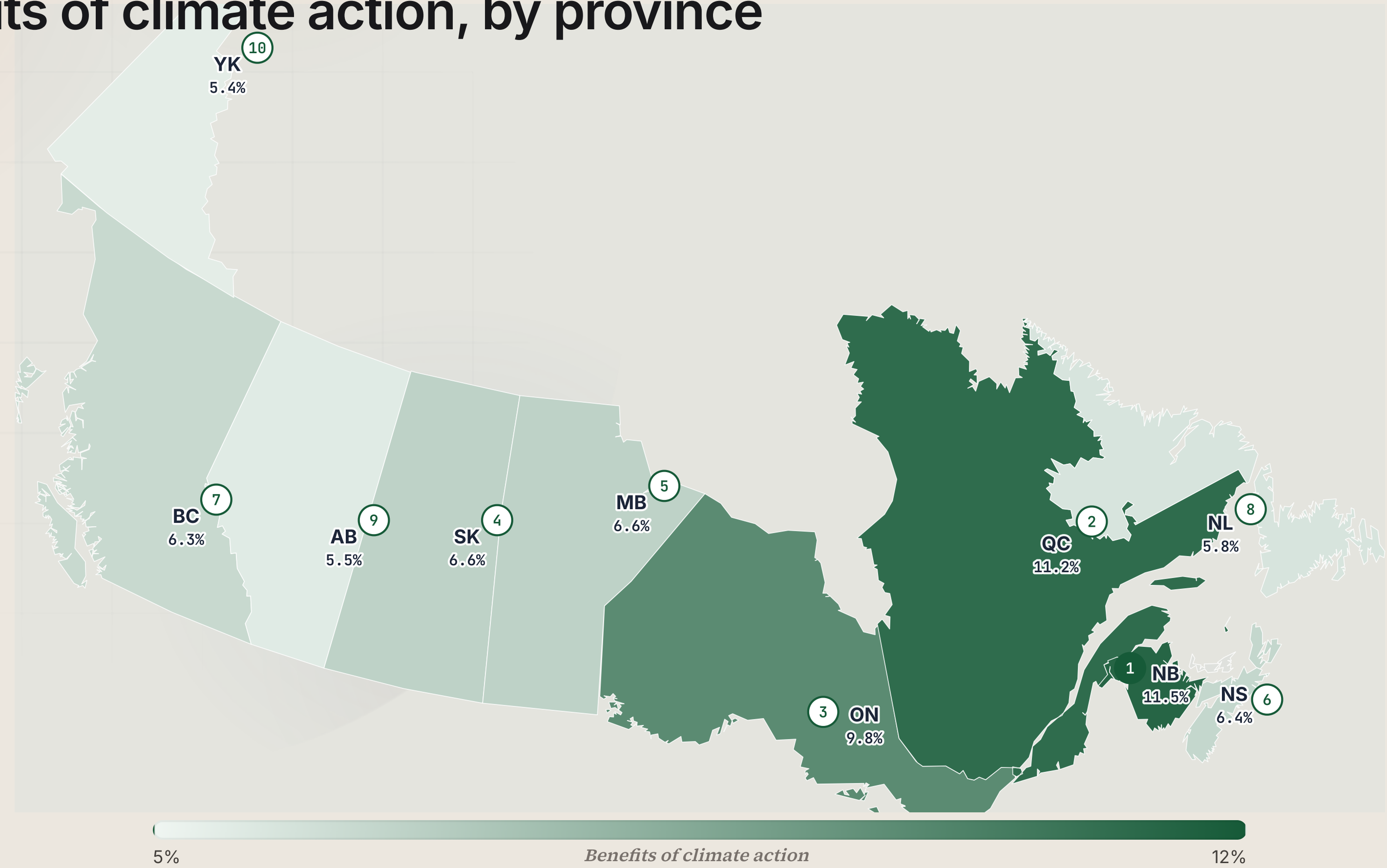
4 MEDIA CASCADES

Can we measure and detect a media cascade, and gauge its effect on framing?

Evolution of thematic frames (1978-2024)

Average sentence share per frame, by year, across the full corpus

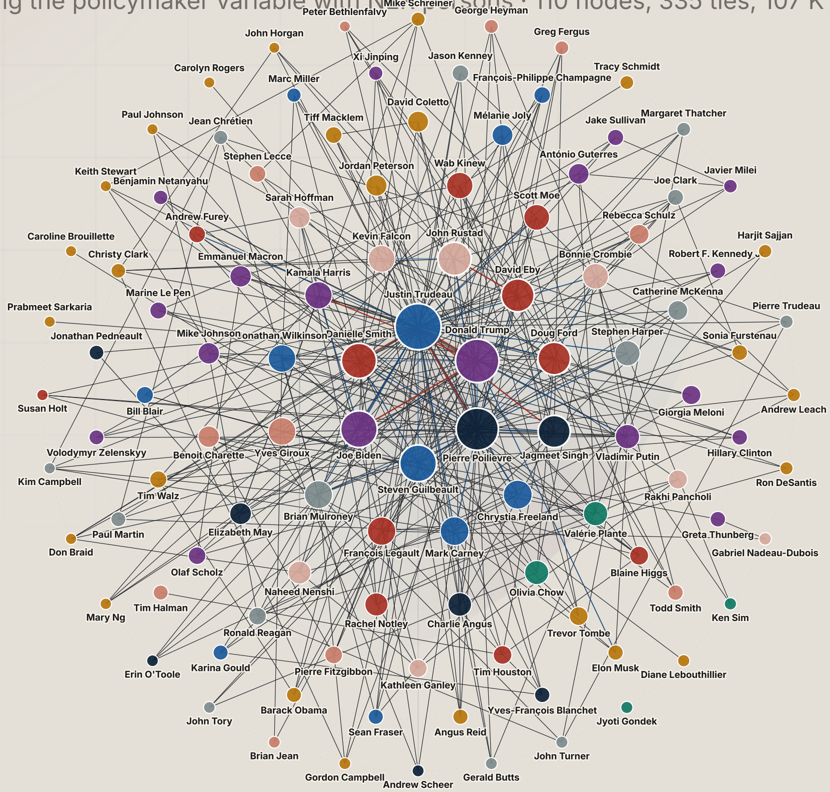
Benefits of climate action, by province



Mean proportion of sentences per article, 2005–2023. Clusters reveal: Quebec → Prairies → Rest of Canada → final ranking.

Who structures climate politics in 2024? · A co-occurrence network

Crossing the policymaker variable with NER persons · 110 nodes, 335 ties, 107 K sentences



110 policymakers · 335 ties

every NER person in 107 K 2024 sentences where the policymaker or political variable = 1

TOP MENTIONS

ROLES

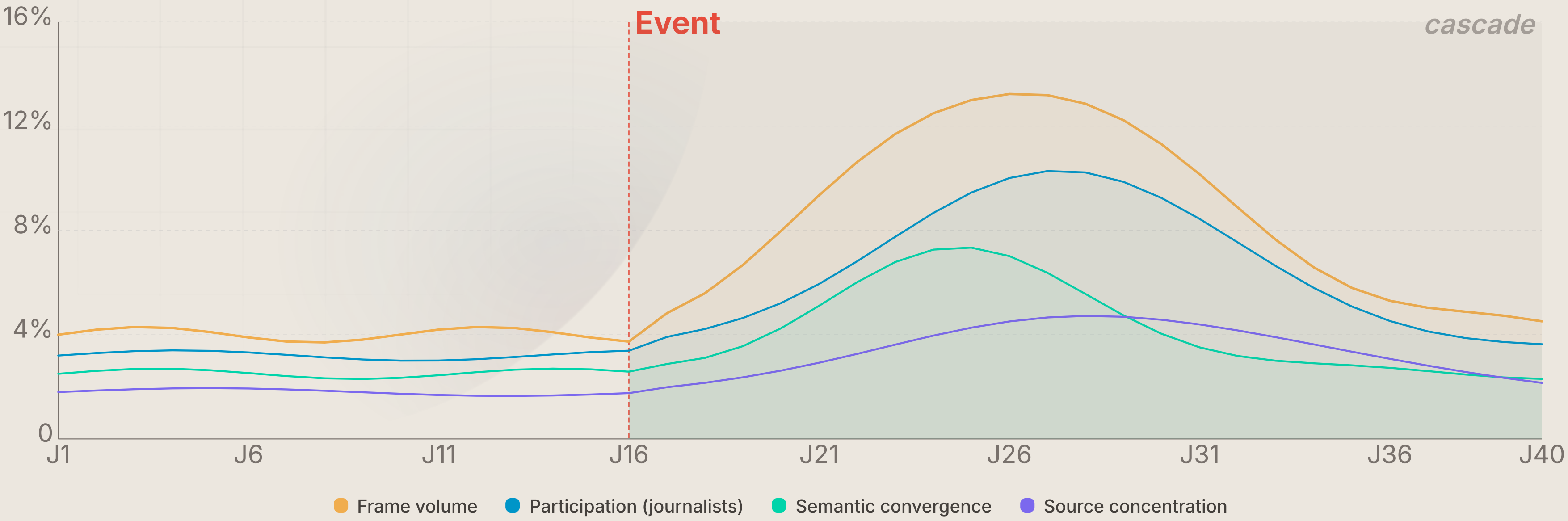
1. 110 nodes · most influential at centre
2. Top 6 names + giant edges (1 300+ shared)
3. Top 14 names + heavy ties (200 to 1 300)
4. Top 35 names + strong ties (50 to 200)
5. Top 55 names + medium ties (10 to 50)
6. Top 85 names + full backbone
7. All 110 names + role legend

What is a media cascade?

A frame propagating massively across outlets after a triggering event

Media cascade =

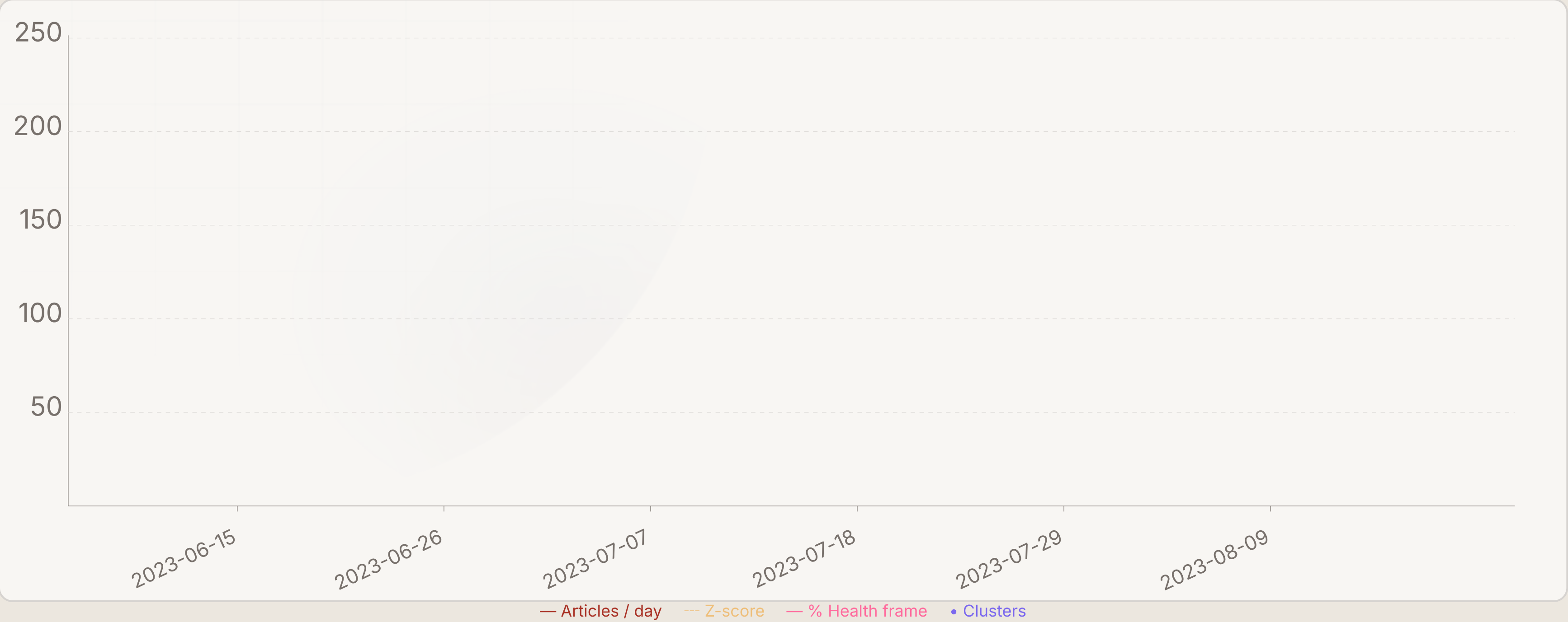
a topic that spreads rapidly across media and journalists, with **convergence** of content, an **increase** in coverage volume, and the arrival of **new actors** in the debate



Health cascade, Summer 2023

0.612 MODERATE CASCADE

453 articles 300 journalists 19 outlets 40 days



PART III

LLM Tool

The puzzle: annotating large corpora without losing the audit trail

Computational social science needs labels for corpora that now contain **hundreds of thousands or millions of texts**.

Grimmer, Roberts & Stewart 2022 · Krippendorff 2018

Zero-shot LLMs make annotation easier, but **model opacity, prompt sensitivity, API drift and costs** weaken cumulative validation.

Weber & Reichardt 2023 · Ollion et al. 2024 · Zhao et al. 2024

Hybrid distillation is promising: LLMs generate training labels, then smaller classifiers reproduce the coding logic across the corpus.

Hinton et al. 2015 · Di Palo et al. 2024 · Pangakis & Wolken 2024

1 **VALIDATION**

Few controlled tests compare LLM-generated training data against human-validated ground truth.

2 **REPRODUCIBILITY**

Published pipelines often depend on undocumented preprocessing, unstable model versions or missing implementation details.

3 **ACCESS**

Without an integrated local workflow, LLM-based annotation remains concentrated among teams with engineering support and API budgets.

RESEARCH QUESTION

Can a locally executable encoder-decoder pipeline produce valid annotations for large corpora while documenting enough of the workflow to make replication and auditing possible?

LLM Tool: a local hybrid workflow, not just another prompt

From a structured codebook to a reusable classifier, with automatic documentation at each stage

1 INPUT

Structured prompt

Codebook, few-shot examples and a dynamic JSON schema (Pydantic) define each task. Text comes in many formats.

JSON SCHEMA · MULTI-FORMAT

→

2 QUALITY CONTROL

Validate + clean

Each LLM response is schema-checked with up to five retries, JSON repair and regex cleaning. Every step is logged.

5 RETRIES · FULL LOGS

→

3 MODEL TRAINING

Fine-tune a classifier

LLM labels train an encoder (50+ models). The best epoch is selected automatically from a combined validation score.

50+ ENCODERS · AUTO-SELECT

→

4 OUTPUT

Annotate the full corpus

The trained model runs high-throughput inference (<1s/doc), parallelized across CPU and GPU, exported in multiple formats.

<1S/DOC · PARALLEL

Every parameter, prompt, model version and environment setting is documented automatically. The pipeline runs entirely on a local machine, no commercial API required, and each run is auditable under documented conditions.

Mode 1 — The Annotator

The Annotator: guided interface for LLM-based classification

MAIN FUNCTION

Uses language models (OpenAI, Ollama) to automatically classify your texts along your category schema.

CAPABILITIES

- Advanced prompt engineering features
- Handles diverse text types
- Runs 100% locally with Ollama (free)
- Auto-repairs format errors
- Exports to CSV, Excel, review platforms

IDEAL USE

Quickly annotate a small corpus, or build training data.

Mode 2 — Annotator Factory

Main function

Chains annotation, data preparation, model training and deployment in a single automated process.

Automated steps

- Annotates a sample with an LLM
- Prepares data (balancing, cleaning)
- Splits train / validation / test
- Selects and deploys the best model
- Applies the model(s) to your whole corpus

Ideal use

A custom model for a large corpus in a few hours.

Annotator Factory: end-to-end model creation

Mode 3 — Training Arena

Training Arena: trains and compares model families

Main function

Trains up to 50 different model types and automatically picks the best one for your data.

Capabilities

— BERT, RoBERTa, DeBERTa, CamemBERT (FR), XLM-R

— Auto-detects the language of your texts

— Compares performance with cross-validation

— Keeps only the best model

— GPU-optimized

Ideal use

Get the optimal model for your training corpus.

Mode 4 — BERT Annotation Studio

MAIN FUNCTION

Applies your trained BERT models to very large document collections at high throughput.

PERFORMANCE

- High-speed GPU annotation
- Auto-optimizes available memory
- Confidence score per prediction
- Handles millions of documents
- Auto-resumes if interrupted

IDEAL USE

Apply your model to large volumes of text.

BERT Annotation Studio: applies trained models on large volumes

Mode 5 — Validation Lab

Validation Lab: agreement metrics and quality control

Main function

Systematically evaluates annotation quality by computing inter-annotator agreement and surfacing problematic cases.

Validation tools

- Cohen's Kappa, Krippendorff's Alpha
- Flags unreliable annotations (score < 0.5)
- Identifies problematic categories
- Stratified samples for human review
- Detailed reports

Ideal use

Ensure scientific quality before publication or final analysis.

Validation design: isolate the human test set

Five LLMs annotate the training pool; five identical XLM-RoBERTa classifiers are scored against human consensus

Canadian political corpus 40 000 sentences · EN / FR · Parliament + media

DISJOINT STRATIFIED SPLIT · ZERO CONTAMINATION

4-TASK CODING SCHEMA

policy themes 21 | political parties 10 | sentiment 3 | specific themes 2

TRAINING PATH 38 451 TEXTS

5 LLM annotators

Gemma 3 · Nemotron · Mixtral · GPT-OSS · GPT-5 (27–120B)

20 XLM-RoBERTa classifiers

5 label sets × 4 tasks · same seed & hyperparameters

GROUND-TRUTH PATH 1 549 TEXTS

4 human experts

CAP codebook · Doccano · independent coding

consensus labels

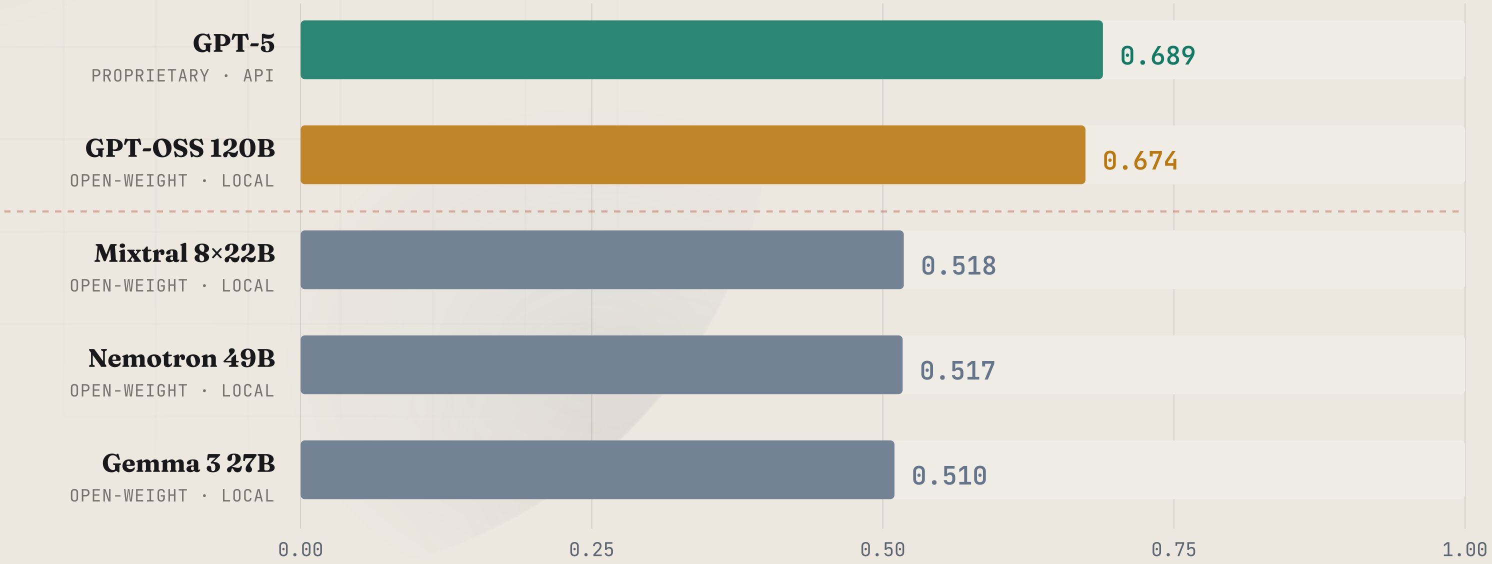
majority vote · $\bar{\alpha} = 0.54$ · $\bar{\kappa} = 0.58$

COMPARATIVE EVALUATION

every classifier scored against the same consensus labels on the isolated test set — micro / macro / weighted F1, with significance tests.

Validation results: two tiers, one local competitor

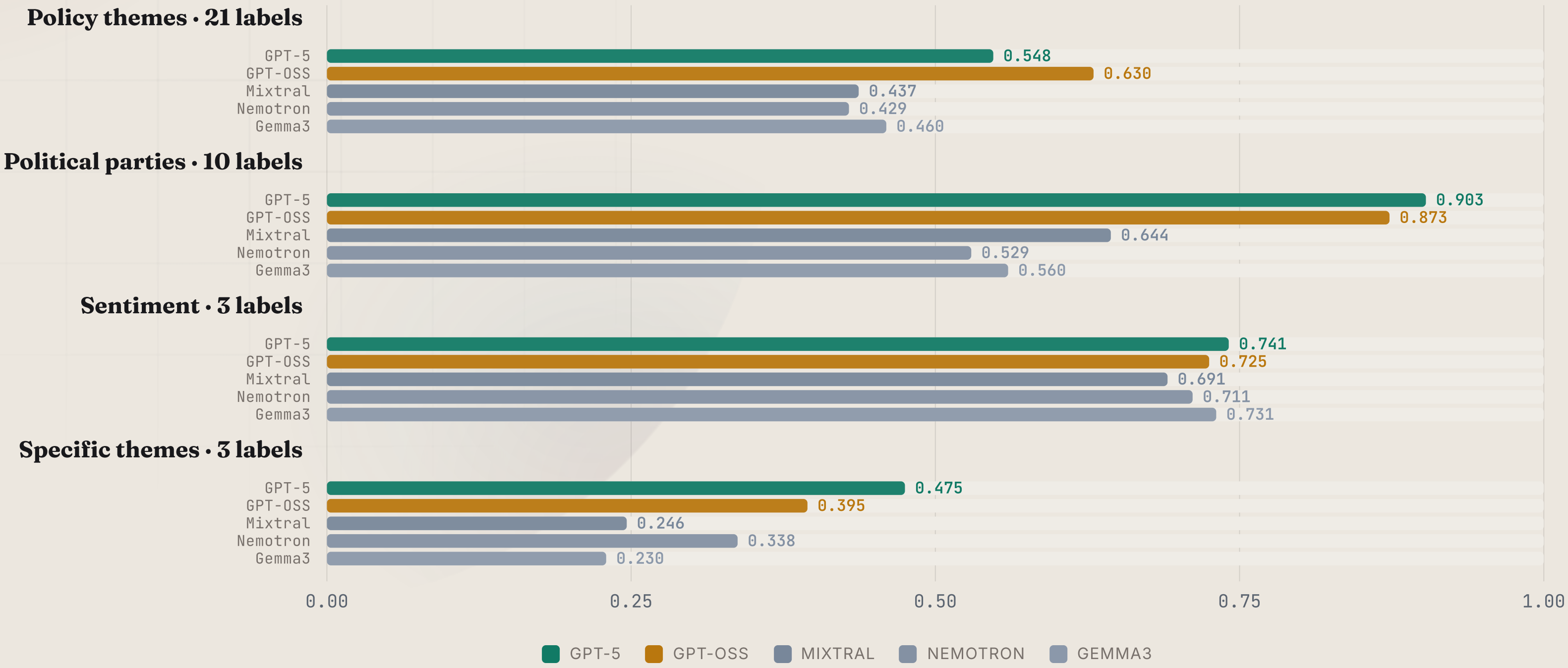
Mean Micro F1 across four annotation tasks, evaluated on the isolated human-consensus test set (N = 1,549)



A 120B open-weight model running locally is statistically indistinguishable from GPT-5 (p = 0.327). The annotator's parameter count does not predict the gap; codebook adherence does.

Validation results by task: Figure 5

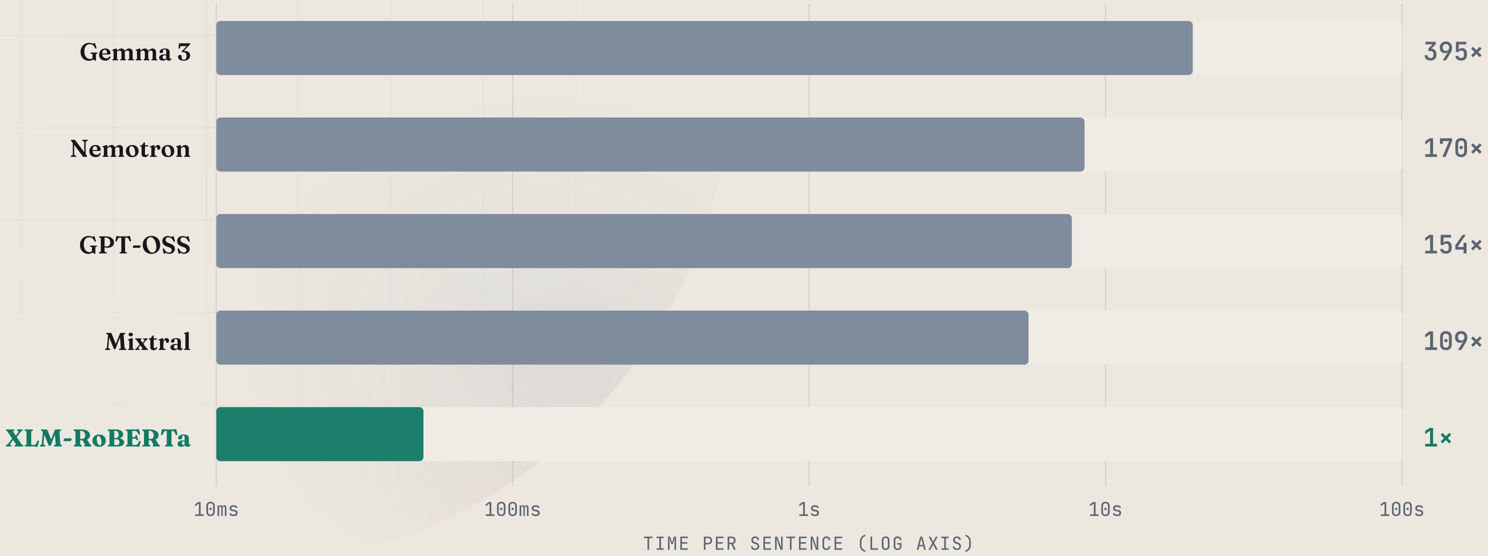
Micro F1 by annotation task and source model, evaluated on the same human-consensus test set



The task matters more than parameter count: GPT-OSS leads on policy themes, GPT-5 leads on parties and sentiment, and specific themes remain difficult because of sparse positive cases.

Validation throughput: 109-395× faster after distillation

Per-sentence inference time after distillation, compared with source-model inference



Source-model inference for 38,451 sentences takes 2.4 to 8.8 days. The distilled classifier processes the same corpus in 32 minutes, a 109× to 395× gain. The initial annotation cost is amortized across every later run.

PART IV

Conclusion

Six takeaways · CCF Database and LLM Tool

1 CCF · A FIRST OF ITS KIND

The first openly released machine-learning-annotated, pan-Canadian climate media database, with 47 years of historical depth and bilingual EN/FR coverage.

2 CCF · QUANTITATIVE OR QUALITATIVE

The same database supports a wide range of research questions, in quantitative and qualitative methods alike, through a web explorer and an open API.

3 CCF · OPEN SCIENCE

Annotations, embeddings and aggregates are openly released on Zenodo, code on GitHub. The data explorer at data.ccf-project.ca is open with code UNIL.

4 LLM TOOL · VALIDATED

The pipeline is evaluated against a human-consensus test set, not assumed valid because an LLM produced labels.

5 LLM TOOL · TRANSPARENT

Prompts, schemas, model versions, hyperparameters, failures and outputs are recorded as part of the method.

6 LLM TOOL · OPEN-SOURCE

The code, data and documentation are available so the workflow can be inspected, adapted and reused beyond climate, for any sentence-level annotation task.

Thank you!

Questions and discussion

Alizée Pillod · Antoine Lemor

Université de Montréal · Université de Sherbrooke

CAPP · CIRST · RFICS

Hosted by Dato Gogishvili and Martin Müller · Team M3 · University of Lausanne · 29 June 2026

ccf-project.ca

PROJECT

data.ccf-project.ca

UNIL

researchsquare.com

PAPER

LLM Tool · GitHub

CODE