CANADIAN POLITICAL SCIENCE ASSOCIATION, TORONTO, JUNE 2-4, 2026

LLM Tool

A hybrid pipeline for automated high-volume text annotation

Local language models and BERT classifiers for computational social science

Antoine Lemor[1], Shannon Dinan[2], Jérémy Gilbert[3]

[1]Université de Sherbrooke · [2]Université Laval · [3]Western University

CAPP · CIRST · RFICS

LLM Tool · GitHub

CPSA2026

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

Why computational?

the annotation bottleneck, the LLM promise and its limits

2

LLM Tool

distillation, validation design, benchmark, modes

3

Applications and results

CCF, YouPol, NRx drift, engagement

PART I

Why computational?

The annotation bottleneck, the LLM promise, and its limits

LLMs in social science: promise and limits

LLMs promise to dissolve the oldest bottleneck in computational social science: annotating corpora large enough to sustain quantitative analysis. They move from a codebook to hundreds of thousands of coded sentences with minimal supervision.

Grimmer et al. 2022 · Gilardi et al. 2023 · Törnberg 2024

But evidence tempers the promise. On complex, multidimensional coding schemes, open models follow detailed codebooks poorly, and **task complexity, not raw capacity, decides whether they succeed**.

Halterman & Keith 2025 · Alizadeh et al. 2024 · Chae & Davidson 2025

THREE CONSTRAINTS

1 **COST**

per-token pricing grows with every new dataset; large corpora become infeasible.

2 **TIME**

local inference on millions of sentences takes weeks to months.

3 **OPACITY**

closed, shifting APIs undermine reproducibility and data privacy.

Ollion et al. 2024 · Weber & Reichardt 2023 · Breznau et al. 2022

THE TENSION

Each approach solves one problem at the price of another: money, time, or transparency. What is missing is infrastructure that combines all three.

The hybrid answer: distillation, and three gaps

Knowledge distillation offers a way through. A large **LLM annotates a few thousand sentences; a compact classifier learns to reproduce that logic** and applies it to millions, at a fraction of the cost.

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

The architecture works in principle, but label noise propagates. Even 80-90% annotation accuracy can bias downstream estimates unless evaluation stays anchored in human-validated ground truth.

Egami et al. 2023 · Nelson 2020

THREE GAPS

1 **VALIDATION**

validation practices are inconsistent; few controlled LLM-vs-human comparisons exist.

2 **REPRODUCIBILITY**

pipelines rarely reproduce: undocumented steps, missing code, unstable model versions.

3 **INTEGRATION**

no integrated, fully local, open-source tool deploys the whole hybrid pipeline end to end.

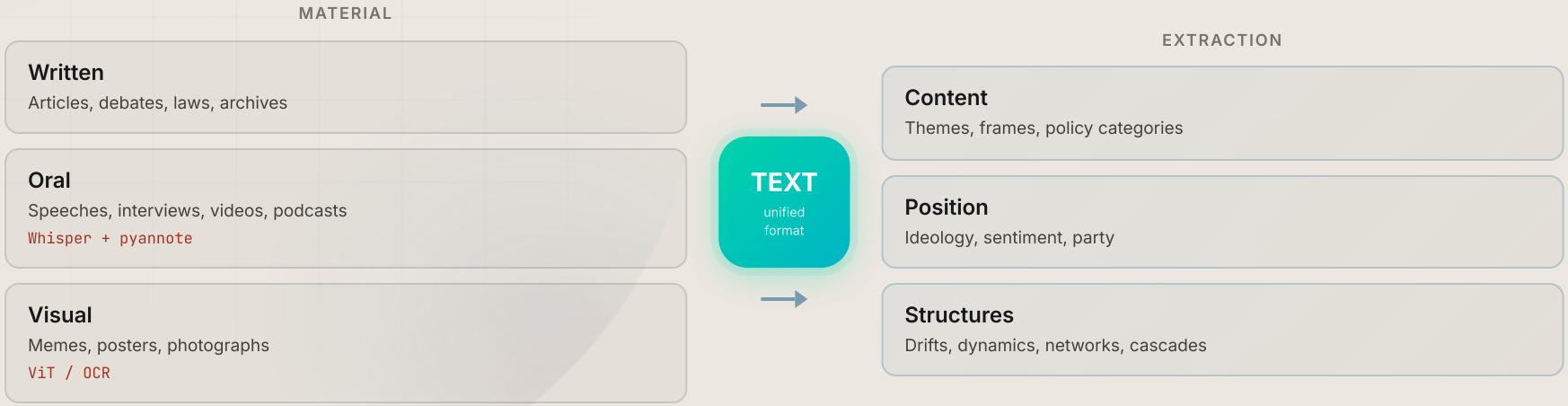
Weber & Reichardt 2023 · Törnberg 2024 · Park & Montgomery 2025

RESEARCH QUESTION

Can a locally executable encoder-decoder pipeline reach near-human annotation quality, and does the choice of LLM annotator still matter once its labels are distilled?

Text: the common format for computational analysis

From diverse materials to a single analyzable format, then to measurable structures



Once everything becomes text, the question is no longer where the data come from, but how to annotate them rigorously across the volumes social science now demands.

What extracting information from sentences makes measurable

Once each sentence carries structured labels, three families of questions open up

1 IDEAS & FRAMES

The progression of ideas

`economic frame` +1.13 pp/decade

Track how an idea or a media frame rises and falls over time. Each sentence becomes one observation in a long time series.

2 ACTORS & COALITIONS

Coalition structure and influence

`actor A → actor B` shared framing

Linking actors to the positions they voice reveals who aligns with whom, which coalitions form, and whose framing others adopt.

3 ACTORS & PRACTICES

Practices tied to actors

`cites evidence` policymaker · 18%

Measure what actors actually do in their speech, for example how often policymakers cite scientific evidence to justify a position.

PART II

LLM Tool

A hybrid pipeline: distillation, validation, and what the benchmark shows

The core idea: knowledge distillation

1 HUMAN EXPERTISE

Codebook + validation

Researchers define the coding scheme and hand-label a small held-out test set. Human judgment stays at the evaluative centre.

GUIDES →

2 TEACHER · LLM

Annotates ~38K sentences

A decoder LLM reads the codebook and produces structured labels. Local (Ollama) or API. Slow but interpretive.

38 451 TRAINING LABELS

TRAINS →

3 STUDENT · BERT

Learns the labeling logic

An encoder classifier (XLM-RoBERTa) reproduces the labels, then annotates the full corpus at a fraction of the cost.

50 ms PER SENTENCE

The quality of a large model at the speed of a small one: one LLM annotates a few thousand sentences so a classifier can annotate millions, while evaluation stays anchored in human judgment.

The pipeline: four phases, fully local

From a structured prompt to a corpus-wide classifier, with automatic documentation at every step

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, fully reproducible.

How we validate the pipeline

Five LLMs annotate the same corpus; their classifiers are scored against human consensus on an isolated test set

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

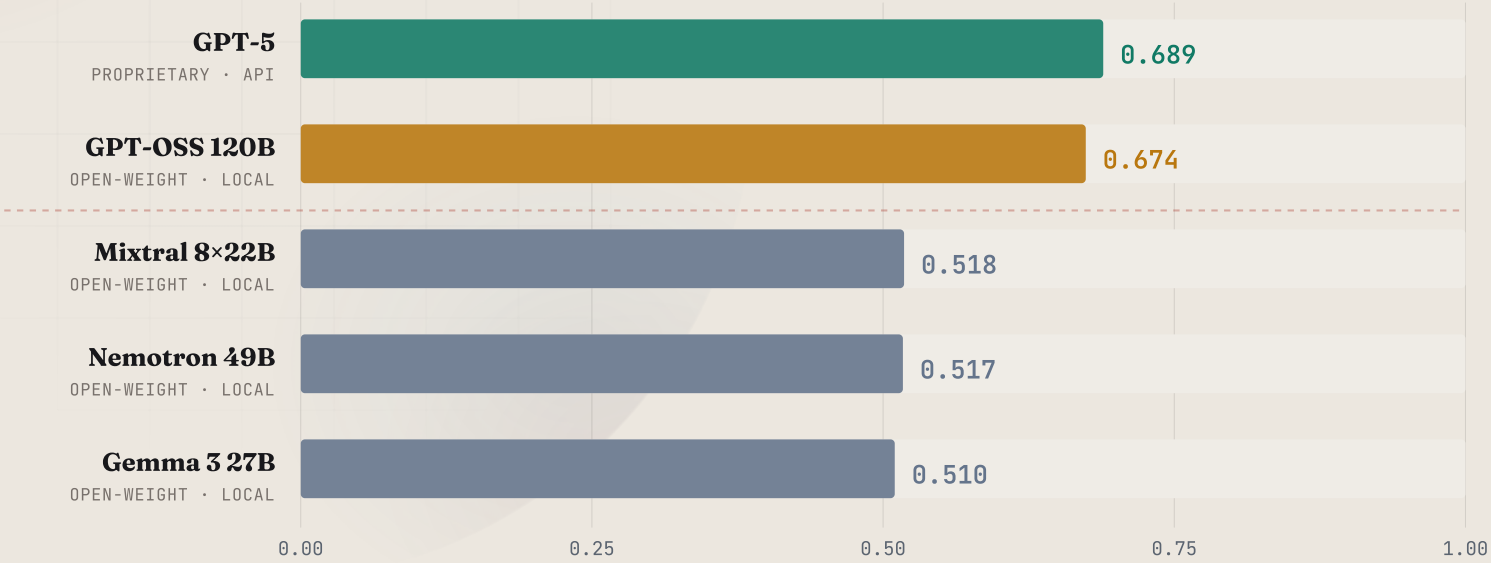
majority vote · $\bar{\alpha}$ = 0.54 · $\bar{\varkappa}$ = 0.58

COMPARATIVE EVALUATION

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

What the benchmark shows: two tiers

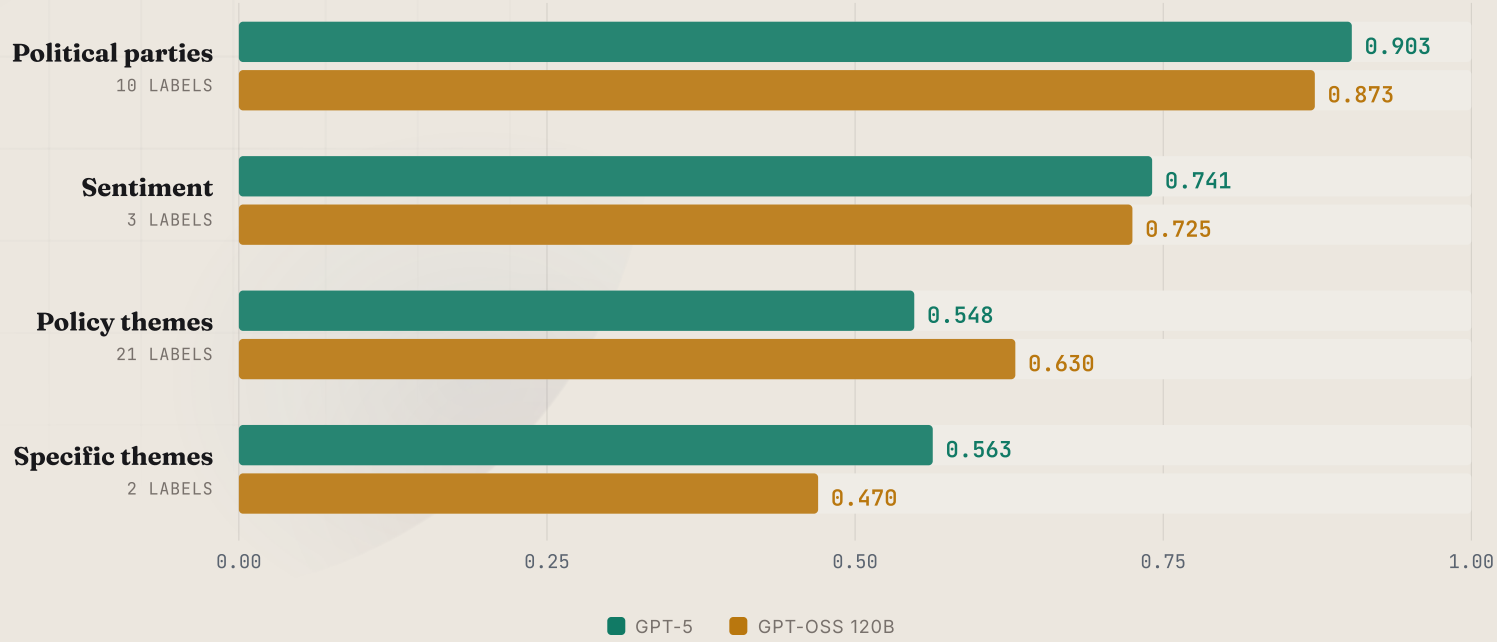
Mean Micro F1 across four annotation tasks, evaluated against human consensus (N = 1 549)



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.

Inside the benchmark: performance by task

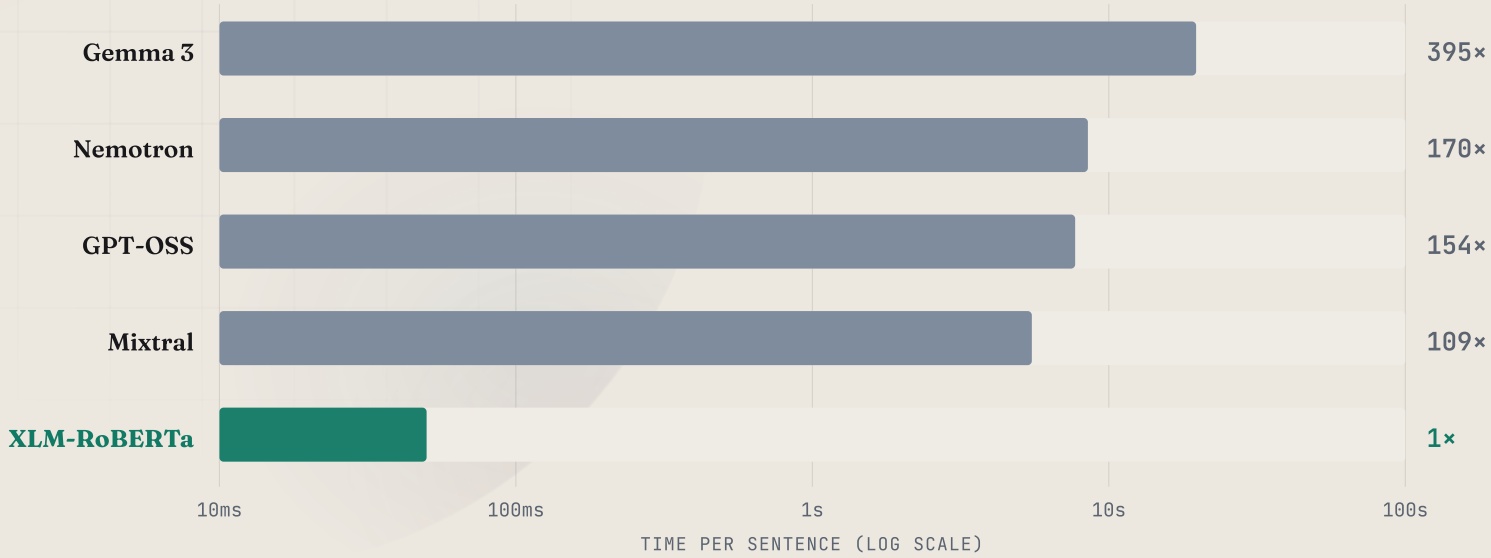
Micro F1 of the two best classifiers across the four annotation tasks, ordered from easiest to hardest



Task complexity drives the gap more than annotator size. On the 21-label policy themes, the open-weight GPT-OSS even overtakes GPT-5 (0.630 vs 0.548); sentiment is easy for both, specific themes hard for both.

What distillation buys: a 109-395× speedup

Per-sentence inference time, LLM annotation versus the distilled classifier



Direct LLM annotation of 38,451 sentences takes 2.4 to 8.8 days. The distilled classifier processes the same corpus in 32 minutes, a 109× to 395× speedup. The one-time annotation cost is amortized across every later run.

Six workflow modes, no code required

1 The Annotator

Zero-shot LLM annotation of a raw text file, exported to Label Studio or Doccano.

2 The Annotator Factory

End-to-end pipeline: chains LLM annotation, data preparation and BERT fine-tuning in one session.

3 Training Arena

Dedicated training and benchmarking workspace: an 18-step wizard from dataset to evaluation.

4 BERT Annotation Studio

High-throughput inference with a trained classifier, parallel GPU/CPU, resumable sessions.

5 Validation Lab

Quality-control workspace: ingests annotator exports and computes Cohen's κ and Krippendorff's α.

6 Resume Center

Session registry: lists partially completed runs so they resume from the last checkpoint.

Every mode is driven from an interactive command line, no code required. Each session persists its configuration, progress and outputs, so any run can be resumed or reproduced later.

PART III

Applications

Two databases built with these methods: measuring frames, detecting drifts

Two databases built with these methods

The rest of this talk shows two concrete applications of the pipeline. Both databases are openly accessible via the QR codes below.

CCF

data.ccf-project.ca

Canadian climate media framing
266K articles · 9.2M sentences · 1978-2024

YouPol

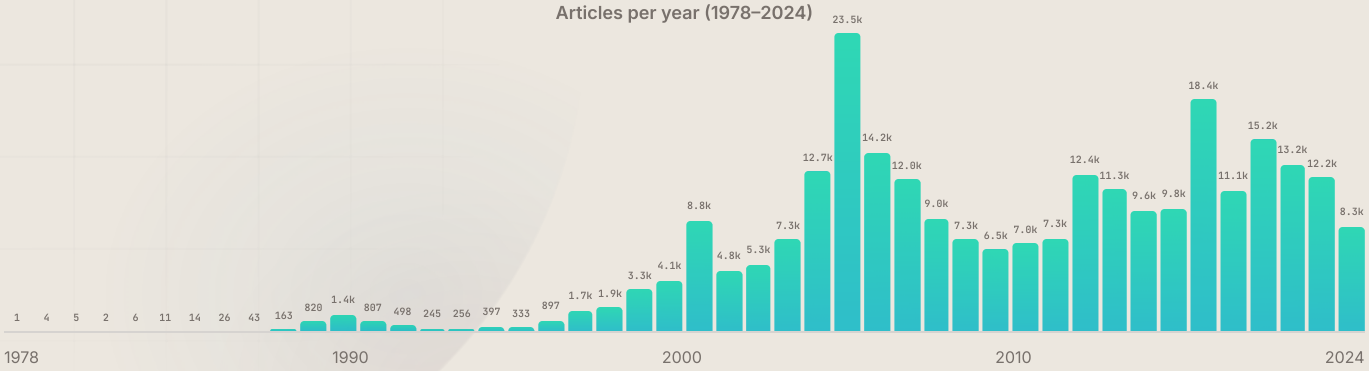
data.you-pol.com

Political video content (YouTube, TikTok)
30K videos · 4.8M sentences · FR + QC

TEMPORARY ACCESS CODE `CPSA2026`

Application 1: the CCF database

266K	20	9.2M	65
Articles	**Newspapers**	**Annotated sentences**	**Categories**
1978 – 2024	EN + FR	BERT / CamemBERT	hierarchical



The largest ML-annotated climate discourse corpus in Canada

Macro F1 : 0.866 – Lemor, Pillod, Taylor (2025)

Evolution of thematic frames (1978-2024)

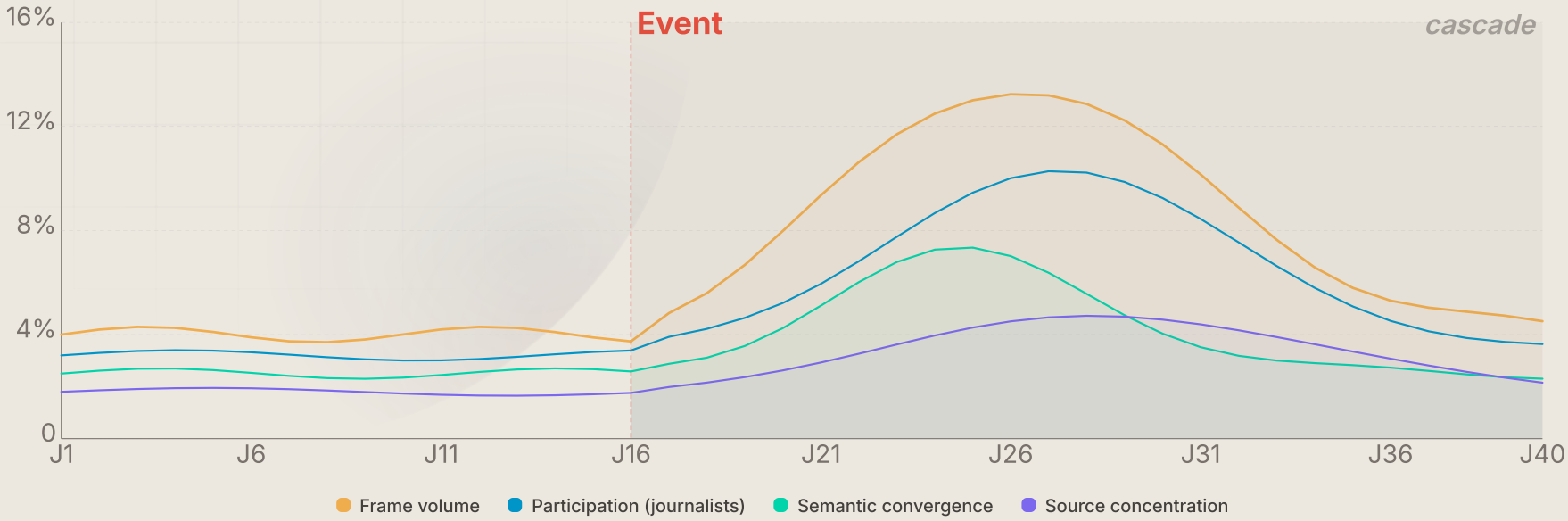
Average proportion of sentences per frame per year

What is a media cascade?

When a media frame propagates massively across outlets

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



Application 2: YouPol

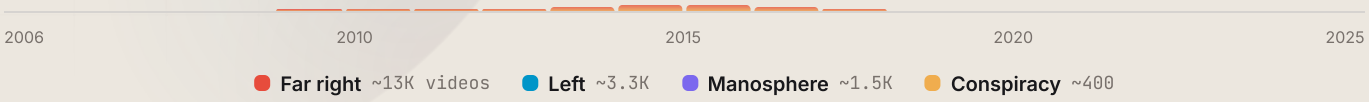
24.7K
Videos
YouTube + TikTok

64
Channels
France + Québec

2.95M
Annotated sentences
CamemBERT

7.6M
Comments
archived

Videos per year (2006 – 2025)



Far right ~13K videos Left ~3.3K Manosphere ~1.5K Conspiracy ~400

Collaborative research infrastructure for political video content
Lemor & Boursier (2026) – you-pol.com – data.you-pol.com

From video to data: the pipeline

Automated steps, from raw video to annotated sentences

1. POLITICAL VIDEO (YOUTUBE / TIKTOK)

"Why mass immigration destroys our civilization"

Far-right channel · 45 min · 180K views

2. AUDIO PROCESSING & TRANSCRIPTION

Demucs → Vocal isolation

pyannote → Diarization

Whisper → Transcription

SaT → Segmentation

3. DIARIZED SEGMENTS

LOC-1	*"La France n'a pas vocation à accueillir la misère du monde..."*
LOC-2	*"...nos ancêtres ont bâti cette civilisation par le sang et le sacrifice..."*
LOC-3	*"...la technologie nous permet maintenant de restaurer l'ordre naturel."*

4. NLP ANNOTATION (LLM TOOL + CAMEMBERT)

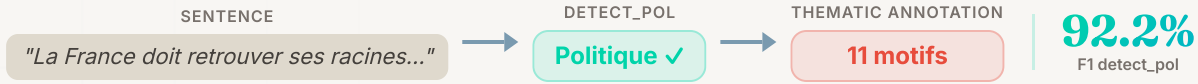
CamemBERTv2 Classifiers trained through knowledge distillation

Political · Immigration · Tradition · Nationalism · NRx Ecology · Technology · Gendered rhetoric · Authority

What YouPol can measure

Thematic motifs across parallel annotation projects

A CamemBERTv2 classifier trained via LLM Tool identifies political sentences before thematic annotation



11 thematic motifs

5 NRx + 6 classical

NRX MOTIFS (NEO-REACTIONARY)

- **Ecology** Reactionary ecology, anti-modernity
- **Equality** Critique of egalitarianism
- **Fictional metaphors** Literary/mythical references
- **Libertarianism** Anti-state, techno-capitalism
- **Technology** Accelerationism, transhumanism

CLASSICAL MOTIFS

- **Authority**
- **Democracy**
- **Immigration**
- **Nationalism**
- **Progress**
- **Tradition**

3 parallel annotation projects

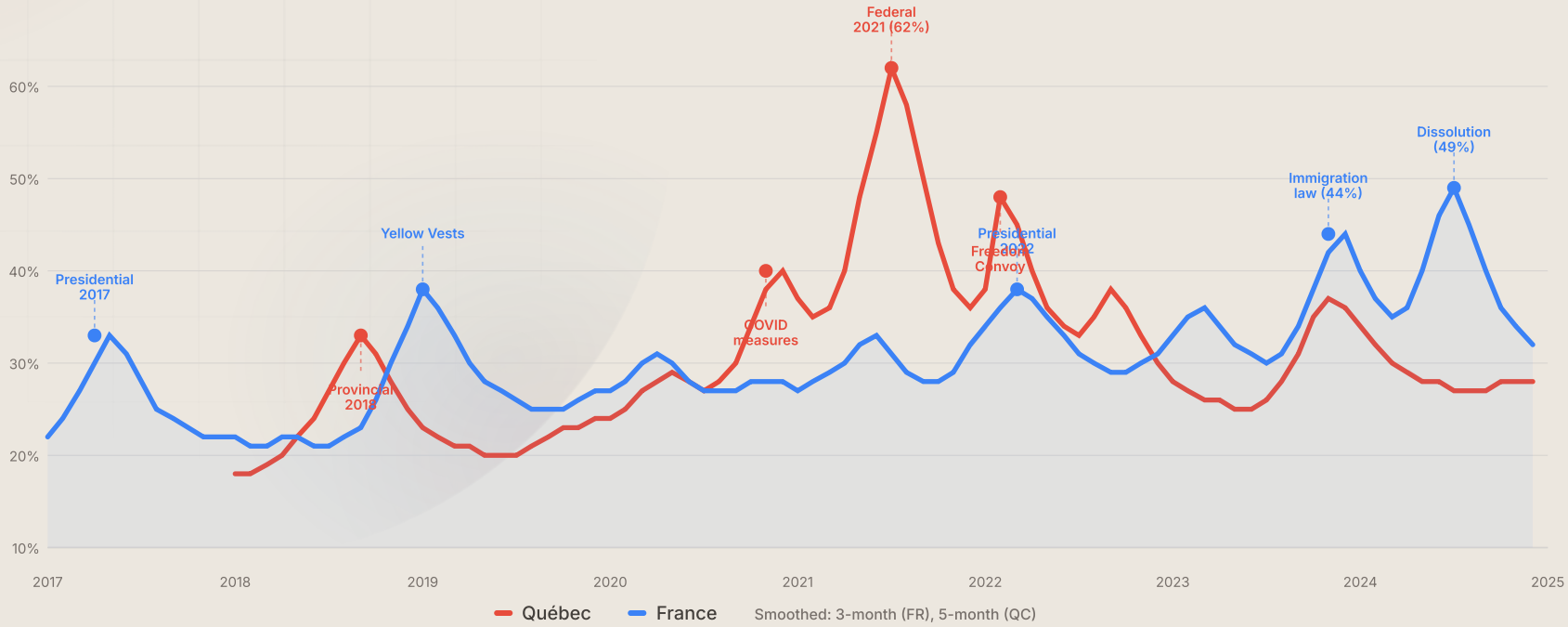
SIED
Socio-Ideological Score
11 macro-categories of far-right discourse

GENRE
Gendered rhetoric
4 dimensions of gendered discourse

NR
Neo-reactionary
Technophilia and reactionary politics

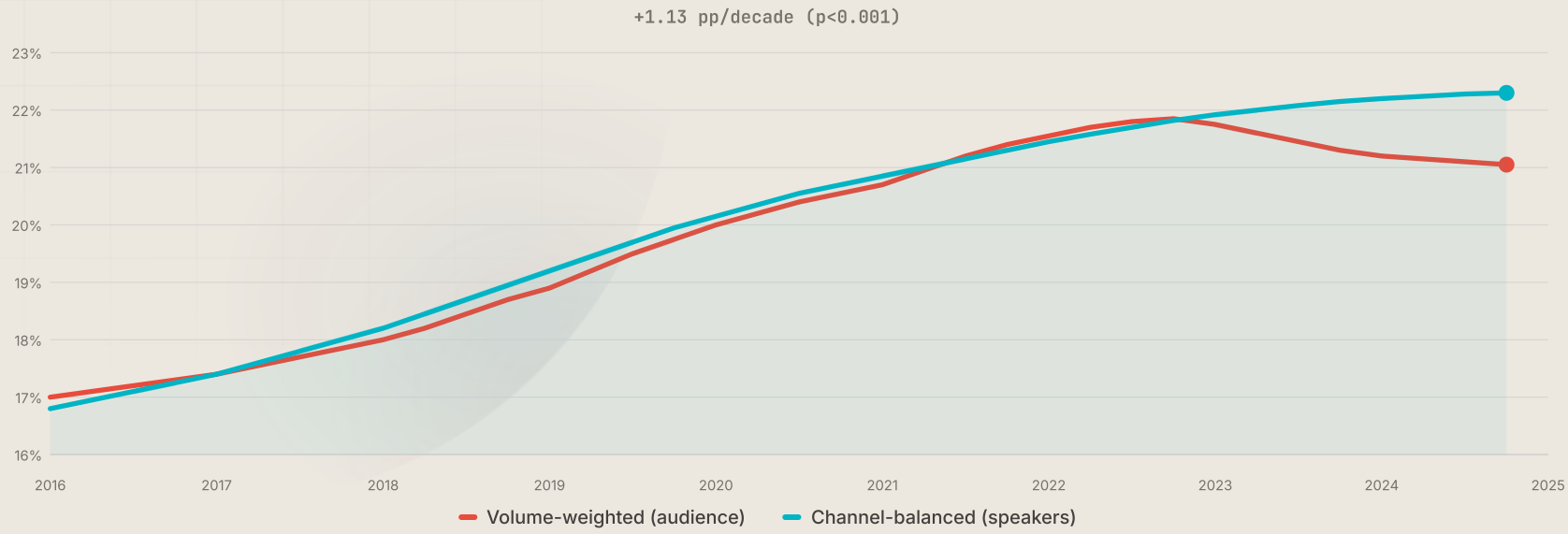
Political density over time (2017-2024)

Monthly share of sentences classified as political by detect_pol, France vs. Québec



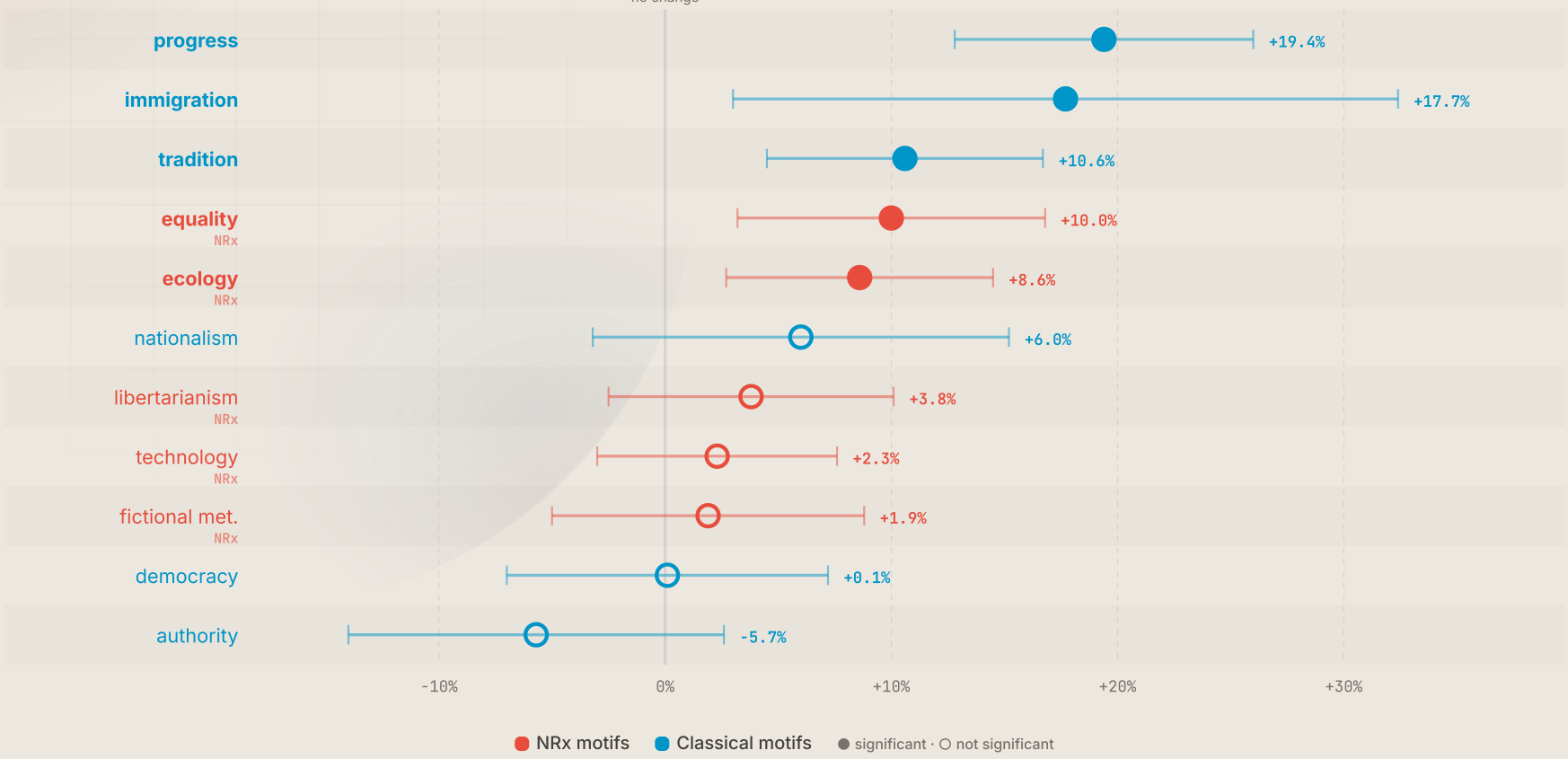
What YouPol shows: NRx drift (2016-2025)

Average NRx motif score across political sentences, stable far-right channels



All themes and political engagement

Effect of each motif on comment politicization (% change per +1 SD, PPML with channel FE)



YouPol: infrastructure and access

1 PRESERVED

Preserved content

2 305
deleted videos

1.58M
comments

Far-right: 11.1% removed · Left: 1.6%

2 LONGITUDINAL

Longitudinal tracking

Full engagement history over time, not a single snapshot.

3 YCCN

YCCN

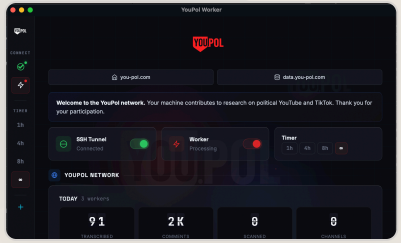
Distributed transcription network. Researchers contribute their own machines.

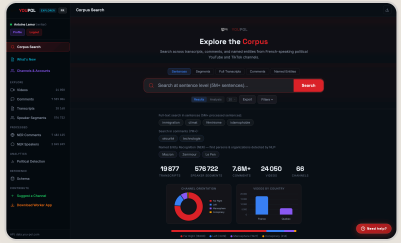
4 API

Open API

data.you-pol.com

REST API, search, export





Thank you!

Questions and discussion

Antoine Lemor · Shannon Dinan · Jérémy Gilbert
Université de Sherbrooke · Université Laval · Western University

LLM Tool (GitHub)

data.ccf-project.ca

data.you-pol.com