

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

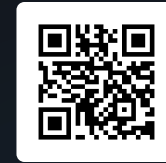
YOUPOL

A database for measuring metapolitics

A continuously updated infrastructure for political video content on YouTube & TikTok

Tristan Boursier¹, Antoine Lemor, Ph.D.²

¹Sciences Po Paris & UQAM, ²Université de Sherbrooke
CIRST, RFICS, CAPP



data.you-pol.com

APSA2026

Outline

1

Why YouPol

political video, metapolitics, four
gaps

2

Building the observatory

channel selection · 7-step pipeline

3

YCCN + Explorer

distributed network, search UI, API

4

Results

scale, political density, deletions

PART I

Why YouPol

Political video, metapolitics, and the gaps in existing data

Video platforms have become central political venues

YouTube hosts an ecosystem of political content creators whose **collective output rivals traditional media** in reach.

RIEDER ET AL. 2020 · MUNGER & PHILLIPS 2022

TikTok has transformed how political content circulates, particularly among younger audiences, through algorithmically formed « *social interest clusters* » where virality matters more than follower counts.

GERBAUDO ET AL. 2024 · GUINAUDEAU ET AL. 2022 · SOLOVEV & PRÖLLOCHS 2026

Francophone political ecosystems remain understudied. Prior work has relied almost exclusively on titles and metadata, overlooking the ideological substance.

BOURSIER 2025 · GILLIOTTE 2024 · RAUCHFLEISCH & KAISER 2020

THE POST-API AGE

CrowdTangle deprecated · TikTok's Research API limited · audits reveal severe biases · YouTube's search API is « *forgetful* », older videos progressively disappear from results.

FREELON 2024 · BRUNS 2019 · RIEDER ET AL. 2025 · BAI 2026

RESEARCH QUESTION

How can we build a research infrastructure that captures what political creators actually say, and that platforms cannot quietly erase?

What is a political influencer? Metapolitics as object

POLITICAL INFLUENCER

A creator whose primary venue is a video platform and whose audience-building strategy fuses entertainment formats (commentary, reaction, interview) with sustained discursive work on political issues.

MUNGER & PHILLIPS 2022 · RIEDER ET AL. 2020 · GERBAUDO ET AL. 2024

METAPOLITICS

A strategy through which far-right actors seek to normalize their ideology by gradually shifting mainstream cultural and discursive frames. Not by chasing electoral wins, but by reframing political issues over time.

BOURSIER 2026 · SCHILK 2025 · GANESH 2025

WHAT THIS MEANS

1 THE FORMAT

Video platforms are not primarily political spaces. Audiences consume the content as entertainment, and ideological arguments arrive embedded in commentary and reaction formats.

2 THE CONTENT

The strategy is only legible through what creators actually say, not through view counts, subscriber metrics, or recommendation graphs.

WHY YouPoL

To study the political strategy and influence of political influencers through what they actually say.

Four gaps in existing political video datasets

DATASET	SOURCE · SCALE	TRANSCRIPTS	DIARIZATION	SENTENCE ANNOTATION	LIVE UPDATES	PRESERVATION	LANG
YouPol	YT + TikTok 24.7K vid · 7.6M comm	<i>full</i>	•	•	•	•	FR + QC
Sosnovik 2025	YouTube auto-captions 100K vid	<i>auto</i>	,	<i>topic</i>	,	,	FR
Pinto 2025	TikTok 3.14M IDs	<i>AI</i>	,	,	,	,	EN
Ribeiro 2020	YouTube metadata 330K vid · 72M comm	,	,	,	,	,	EN
ParlSpeech V2	Parliament 6.3M speeches	<i>official</i>	,	,	,	,	9 lang
Media Cloud	News articles 2B stories	<i>text</i>	,	,	•	,	20+

1 GAP

Metadata cannot capture what creators actually say, so the ideological substance of political video stays empirically out of reach.

2 GAP

Once a video or a channel is removed by the platform, the material is gone for researchers who had not anticipated the deletion.

3 GAP

Single-point snapshots of metadata miss how political attention develops over weeks and months around an event.

4 GAP

No existing dataset preserves comments over time, so moderation as a political communication strategy remains unobservable.

A March 2025 deletion that disappeared 1 755 videos

The FDENEWS case, what survivorship bias looks like, and why pre-emptive collection matters

EVENT

12 March 2025

YouTube permanently deletes FDENEWS for repeated community-guideline violations.

WHAT DISAPPEARED FROM THE PLATFORM

1,544

VIDEOS

52,774,695

CUMULATIVE VIEWS

224,670

COMMENTS

87,398

SUBSCRIBERS

WITHOUT YouPol

Thousands of hours of political commentary, hundreds of thousands of user comments, and a longitudinal record of one of France's most prominent far-right outlets simply ceased to exist.

WITH YouPol

Every video already downloaded. Every transcript already produced. Every comment already archived. The deletion itself is logged as a data point.

The content researchers most need to study is precisely the content most likely to be removed. Without pre-emptive collection, we study only what platforms allow to remain.

PART II

Building the observatory

Channel selection · 7-step pipeline · LLM-in-the-loop annotation

Three design principles

1 PRINCIPLE PERMANENCE

Continuous monitoring captures content before platforms can remove it. Deleted videos and suppressed channels are flagged but preserved in full.

2 PRINCIPLE DEPTH

Full transcriptions with speaker diarization enable sentence-level analysis of what creators actually say, rather than inference from titles or metadata.

3 PRINCIPLE COLLABORATIVE

The YouPol Collaborative Computing Network lets any researcher contribute processing capacity from their own machine. No institutional cluster required.

The pipeline is language-agnostic. Anglophone expansion only requires new channel identification and adapted codebooks.

Step 1, Channel selection: top-down + snowball

TOP-DOWN

Channels are seeded from prior academic mapping, then enriched through audience metrics (views, subscribers) and media watchdog reports.

RAUCHFLEISCH & KAISER 2020 · MUNGER 2022 · BOURSIER 2025

SNOWBALL

An iterative protocol follows YouTube's related-channel and recommendation graph until saturation, adding channels invisible to audience metrics.

REVEILHAC ET AL. 2024 · GILLIOTTE 2024 · LEWIS 2018

INDEPENDENT VALIDATION

The corpus is cross-checked against Gilliotte (2024), an independent mapping of 127 French political YouTube channels. Coverage of the major clusters is confirmed.

1

AXIS · ORIENTATION

far right left manosphere conspiracy

2

AXIS · COUNTRY

FRANCE QUEBEC

3

AXIS · GENDER

MEN WOMEN MIXED

The current corpus, 30 062 videos across 69 channels

24,164

VIDEOS

63

CHANNELS

2

COUNTRIES

4

ORIENTATIONS

72.8%

19.8%

6.0%

Far right

35 CHANNELS

FR 14,636

QC 4,156

18,186

Left

13 CHANNELS

FR 5,098

4,934

Manosphere

11 CHANNELS

FR 647

QC 907

1,504

Conspiracy

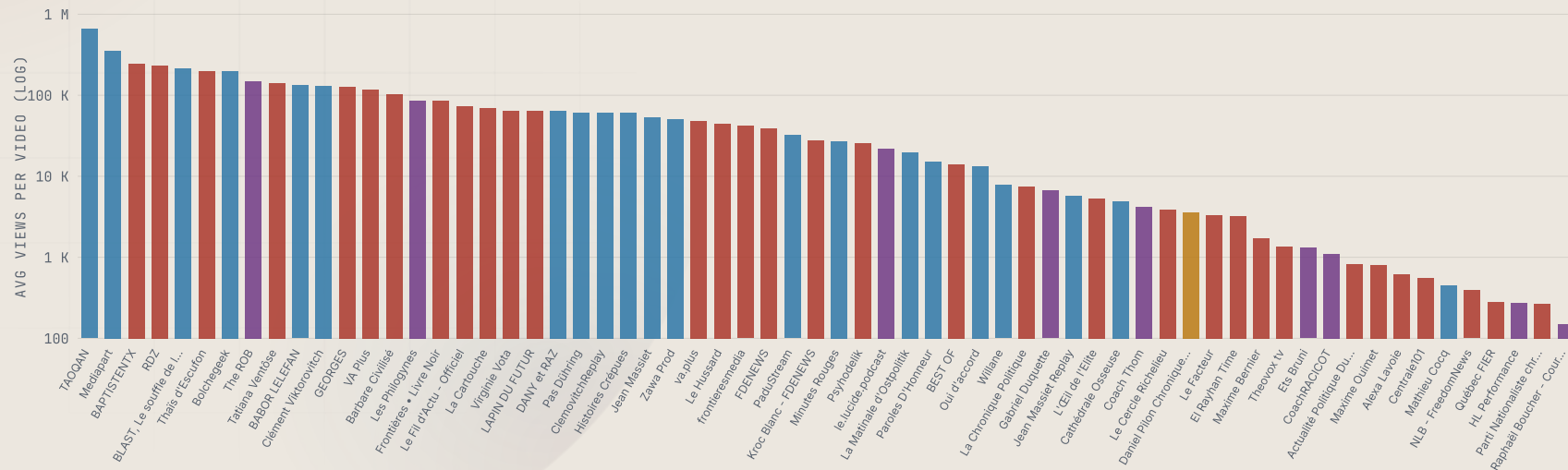
9 CHANNELS

QC 355

344

LIVE COUNTS · PRODUCTION DATABASE · MAY 30, 2026

All 64 channels, ranked by reach per video

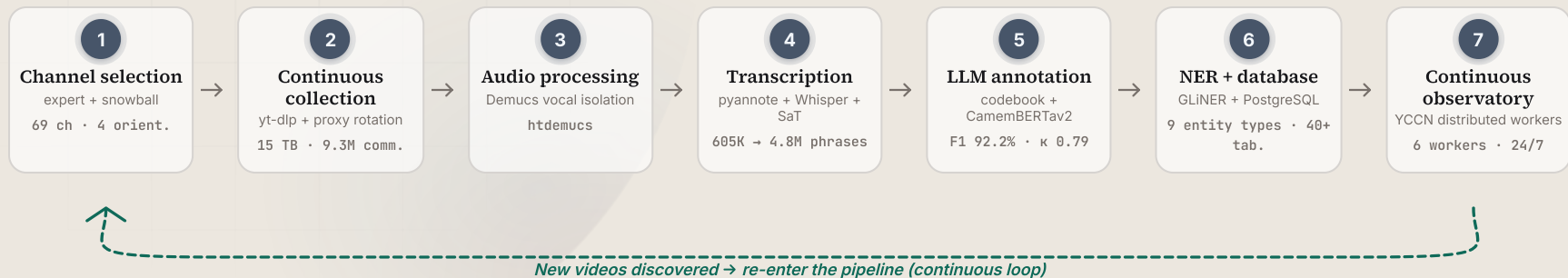


Far right Left Manosphere Conspiracy

0 CHANNELS RANKED BY REACH PER VIDEO

The seven-step processing pipeline

The observatory operates as a continuous loop: step 7 feeds new videos back into step 2



Steps 2-4, From raw audio to diarized sentences

1 raw vocal track plus music and ambient



2 Demucs , *htdemucs* source separation

VOCAL STEM ISOLATED



3 pyannote.audio , *speaker diarization*, threshold 0.75



4 Whisper large-v3 + SaT , *WER 5.6%*, *sentence boundaries*

SPK_00 "Today we will talk about the immigration question."

SPK_01 "Yes, and it is a key issue for the next election."

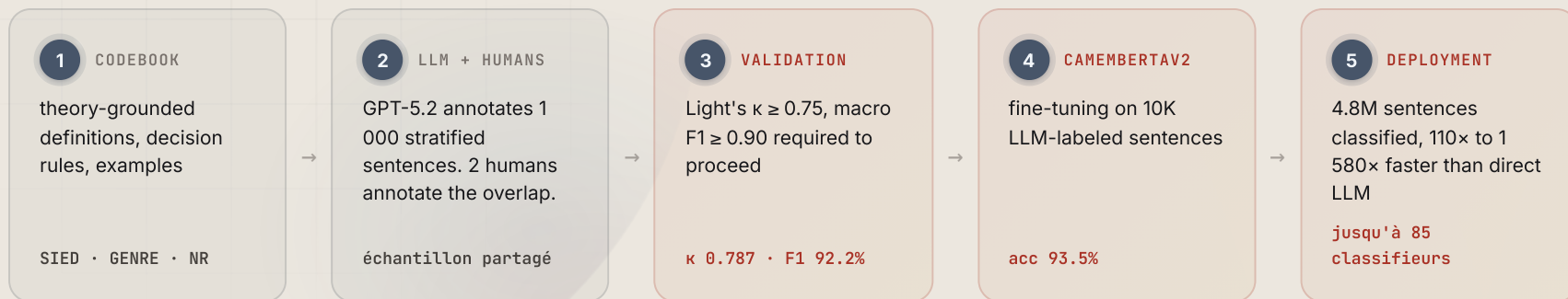
454,302 → 3,175,336
SPEAKER SEGMENTS INDIVIDUAL SENTENCES

Step 5, LLM-in-the-loop annotation with CamemBERTav2



LLM_Tool

Knowledge distillation: LLM labels (κ 0.787) train lightweight classifiers (F1 92.2 %) for corpus-wide deployment



↻ uncertain predictions feed back into the LLM sample, then into a new training round

ANNOTATION PROJECTS IN PRODUCTION

detect_pol
binary political gate

DEPLOYED

SIED
IN ITERATION
far-right ideology · 11 dimensions

GENRE
IN ITERATION
gendered rhetoric · 4 dimensions

NR
IN ITERATION
neo-reactionary technophilia

Step 6, Named entities + PostgreSQL integration

GLINER · ZERO-SHOT EXTRACTION · 9 ENTITY TYPES

« Macron **PERSON** announced a 100-billion recovery plan with the support of the RN, **POLITICAL PARTY** said the daily Le Monde. **MEDIA** »

Entity types extracted



queryable by type, name, or co-occurrence

PostgreSQL — 40+ tables, 5 groups

Content	videos · channels · comments	30 062 · 69 · 9.3M
Transcripts	segments · sentences · processed	1.02M · 4.8M · 4.8M
Metadata history	video_meta · channel_meta	33 746 · 620
Pipeline	workers · new_videos · events	6 · 16 167 · 14K
Combined views	videos (YT+TT) · speakers_with_pol	unified API

PostgREST · JWT · role-based · audit log

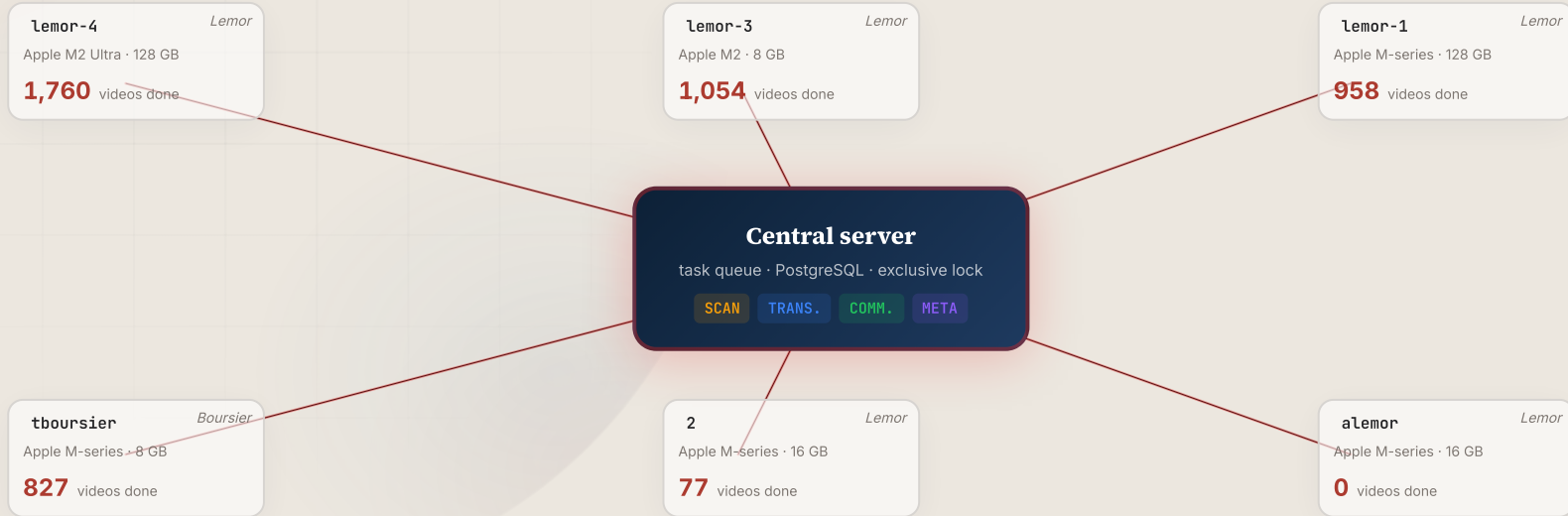
PART III

YCCN + Data Explorer

Step 7, distributed computing · search interface · open API

Step 7, The YouPol Collaborative Computing Network

*A central server creates tasks and places them in a shared queue. Each worker autonomously claims, processes, and releases tasks.
Compute capacity grows as new collaborators join.*



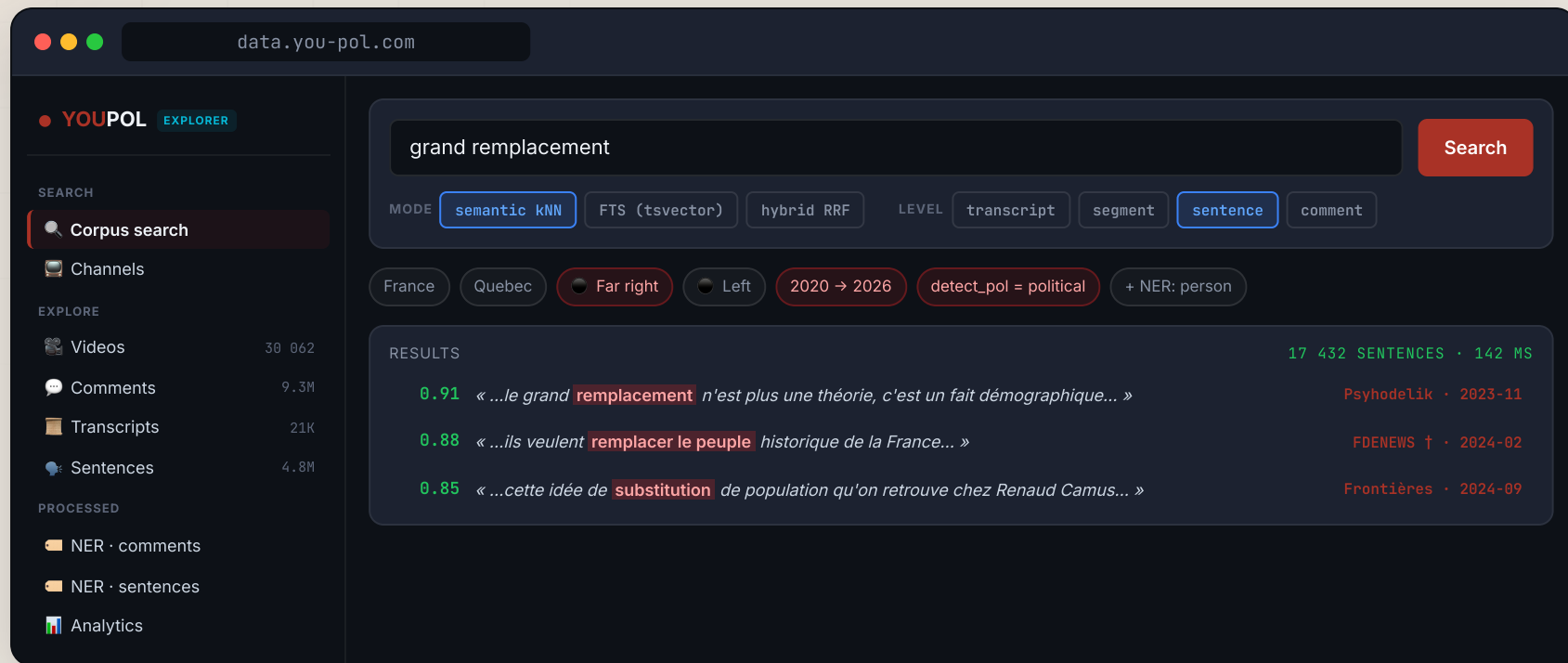
3,964

VIDEOS PROCESSED BY COLLABORATORS

No institutional cluster required. The infrastructure scales with its community.

The Data Explorer interface

A live mockup of data.you-pol.com, corpus search · channels · analytics



Three retrieval modes · four corpus levels · NER + ideology + temporal filters · CSV / JSON export · Python client

Added since the paper

Four capabilities built between the JCSS submission and CPSA, embeddings, augmentation, sampling, daemon

CAPABILITIES ADDED BETWEEN THE JCSS SUBMISSION AND THIS CONFERENCE

1 Semantic embeddings

3,805,061
SENTENCES ENCODED

Qwen3-Embedding-8B · 4 096-d halfvec

Sentences, segments and transcripts encoded into half-precision vectors.
Powers semantic kNN search and the reranker on the explorer.

2 LLM snowball augmentation

ITERATIVE, IN-CORPUS

Ollama · Qwen3 · iterative

A researcher seeds a few examples, an LLM proposes new candidates from the corpus, the feedback loop builds a balanced training set.

None of these change the paper's claims. They sharpen the tools researchers actually reach for.

Three modes of access

1 WEB EXPLORER

Interactive search across transcripts, segments, sentences and comments. Visual filters by channel, orientation, country, date and NER entity.

`data.you-pol.com`

2 REST API

PostgREST endpoint with JWT authentication. SQL-like filtering, pagination, bulk CSV / JSON exports, role-based access with audit logging.

`PostgREST · 50+ endpoints`

3 PYTHON CLIENT

A wrapper for the API. Retrieve full transcripts or comment threads, export to analysis workflows, for both quantitative and qualitative research.

`pip install youpol-client`

4 ACCESS DURING THIS CONFERENCE



`data.you-pol.com`

`APSA2026`

PART IV

Results

Scale · political density · temporal dynamics · preserved content

The corpus at a glance

21,127

videos · YT + TikTok

+ 16 167 in pipeline

3,371,754

annotated sentences

speaker-diarized · sentence-level

6,548,956

archived comments

YouTube + TikTok (234 811)

1,162,075,281

cumulative views

reach captured by the corpus

55 CHANNELS

2 PLATFORMS

27,016 METADATA SNAPSHOTS

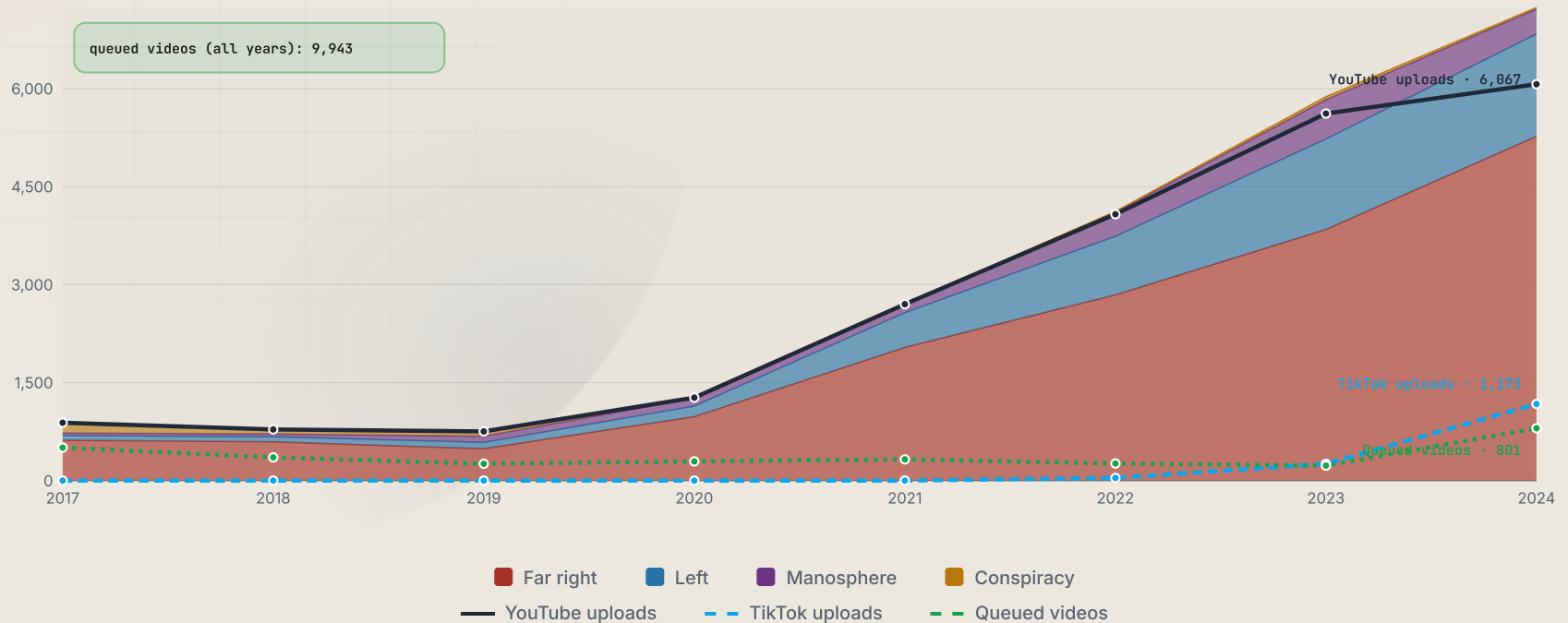
12 TB AUDIO PROCESSED

1,845 PRESERVED DELETIONS

LIVE COUNTS · PRODUCTION DATABASE · MAY 30, 2026 · PAPER FIGURES REFLECT THE APRIL 2026 SNAPSHOT

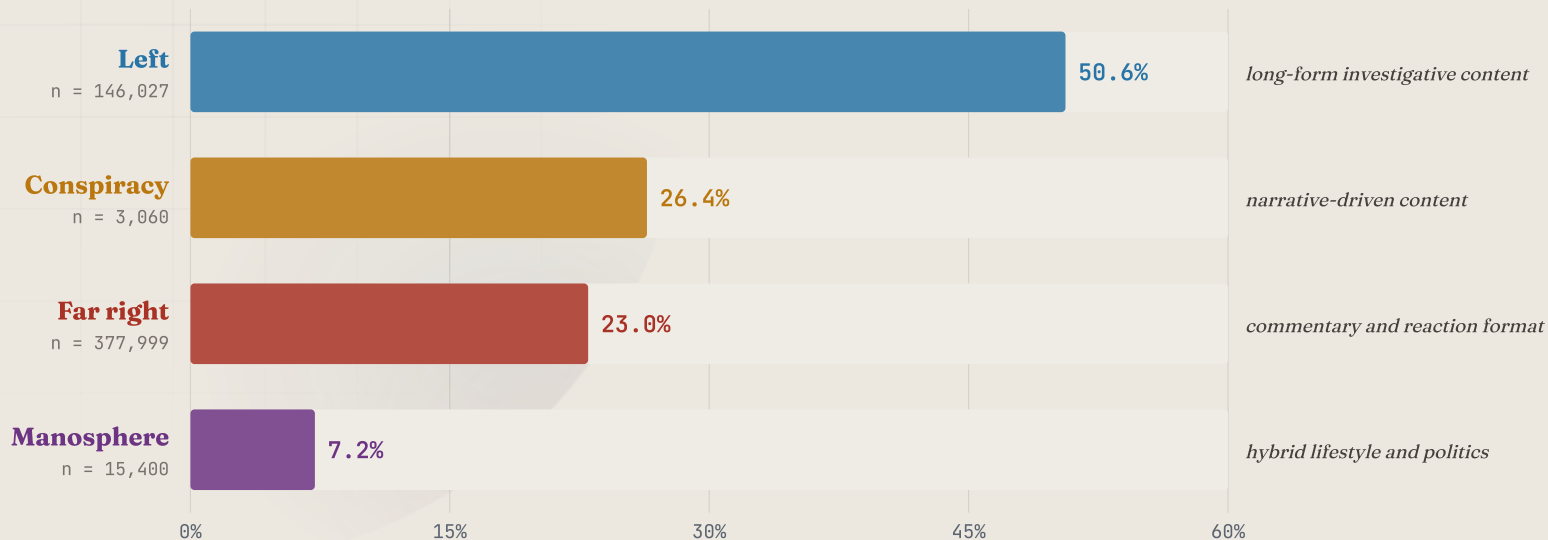
Temporal growth of the corpus (2017-2024)

Stacked ideological uploads by year, with YouTube / TikTok / queued overlays from the production server



Political density by ideological orientation

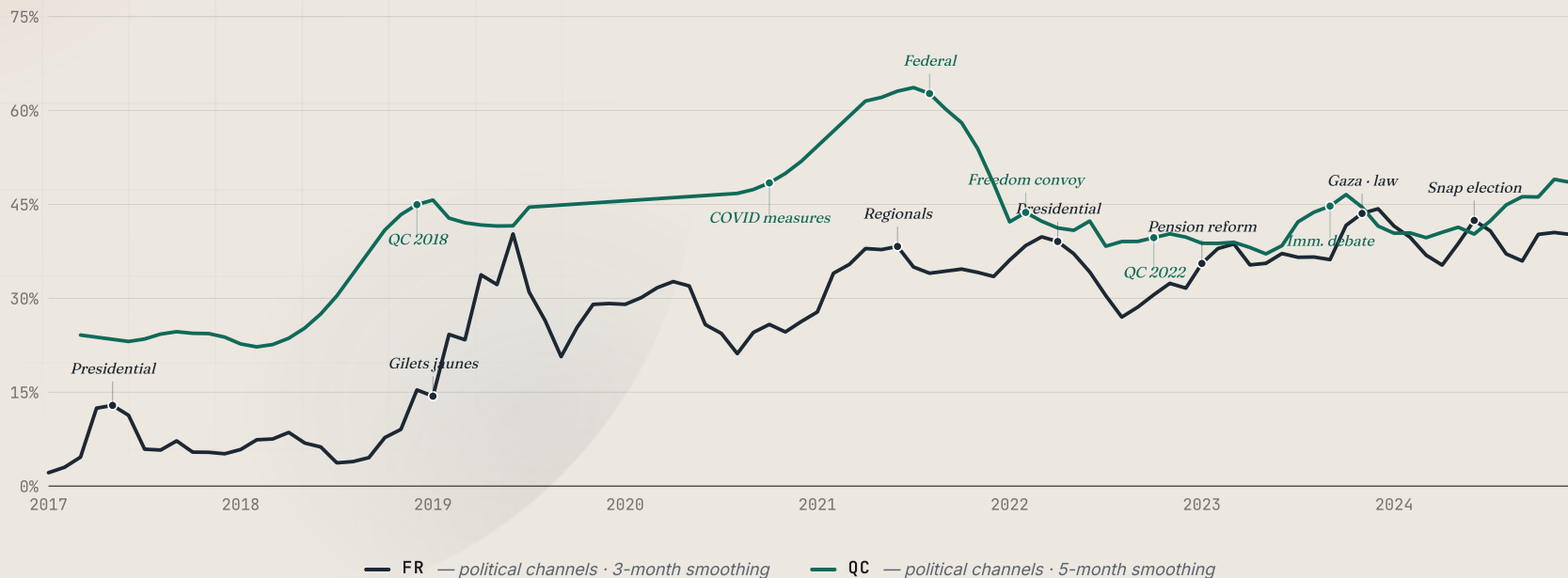
detect_pol classifier deployed on 542K sentences from the political density sample



Political density spans a 7x range. It is a property of the format as much as of the ideology.

Two distinct national calendars: France vs Quebec

Monthly political density · each ecosystem responds to its own political events, not to a shared dynamic

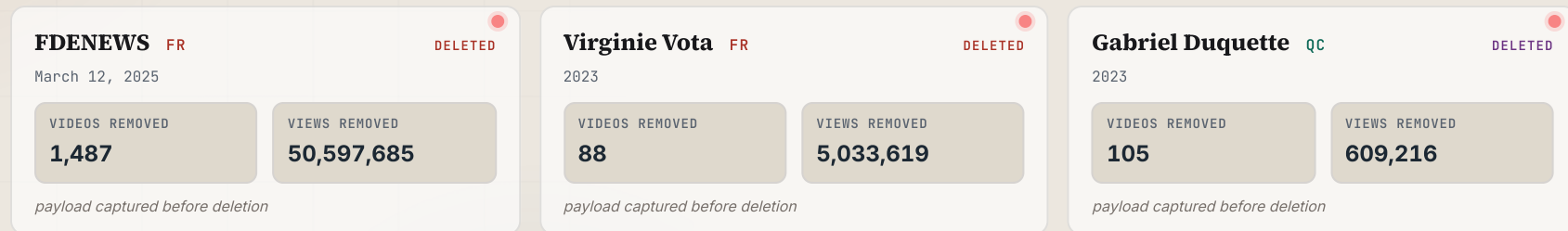


METHOD NOTE

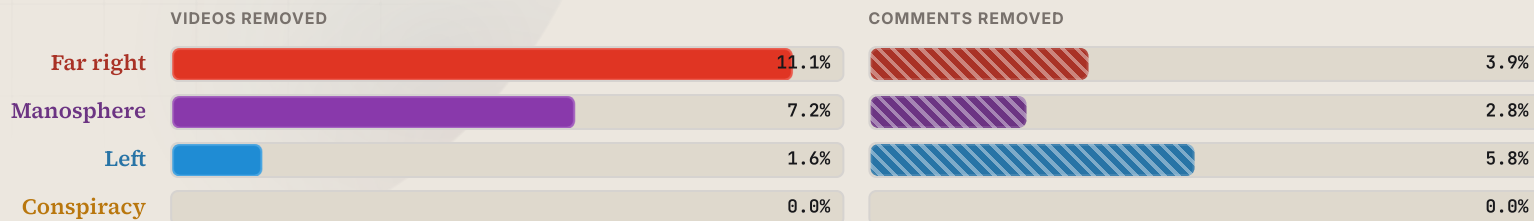
The political-density score aggregates channels classified as far-right, left, or conspiracy. Curves are smoothed (3 months for FR, 5 months for QC) to emphasize structural national cycles rather than month-level noise.

Preserved content, 2 305 deletions + 3 channels

† Three channels entirely deleted by YouTube — preserved in full



Suppression patterns diverge by orientation — itself an analytic finding

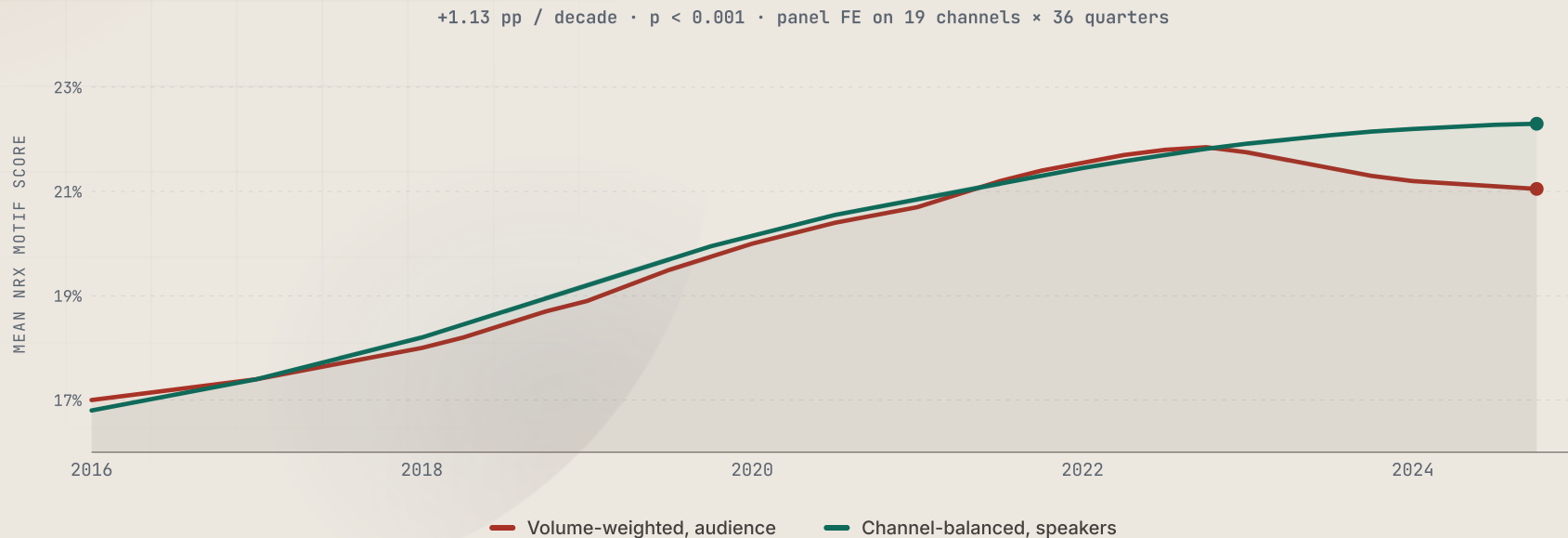


PRESERVATION NOTE

Without continuous monitoring, **1,953** videos and **1,168,428** comments would be permanently lost to research.

Detecting a neo-reactionary drift, 2016 to 2024

Mean NRx motif score on 335K political sentences from 19 stable far-right channels

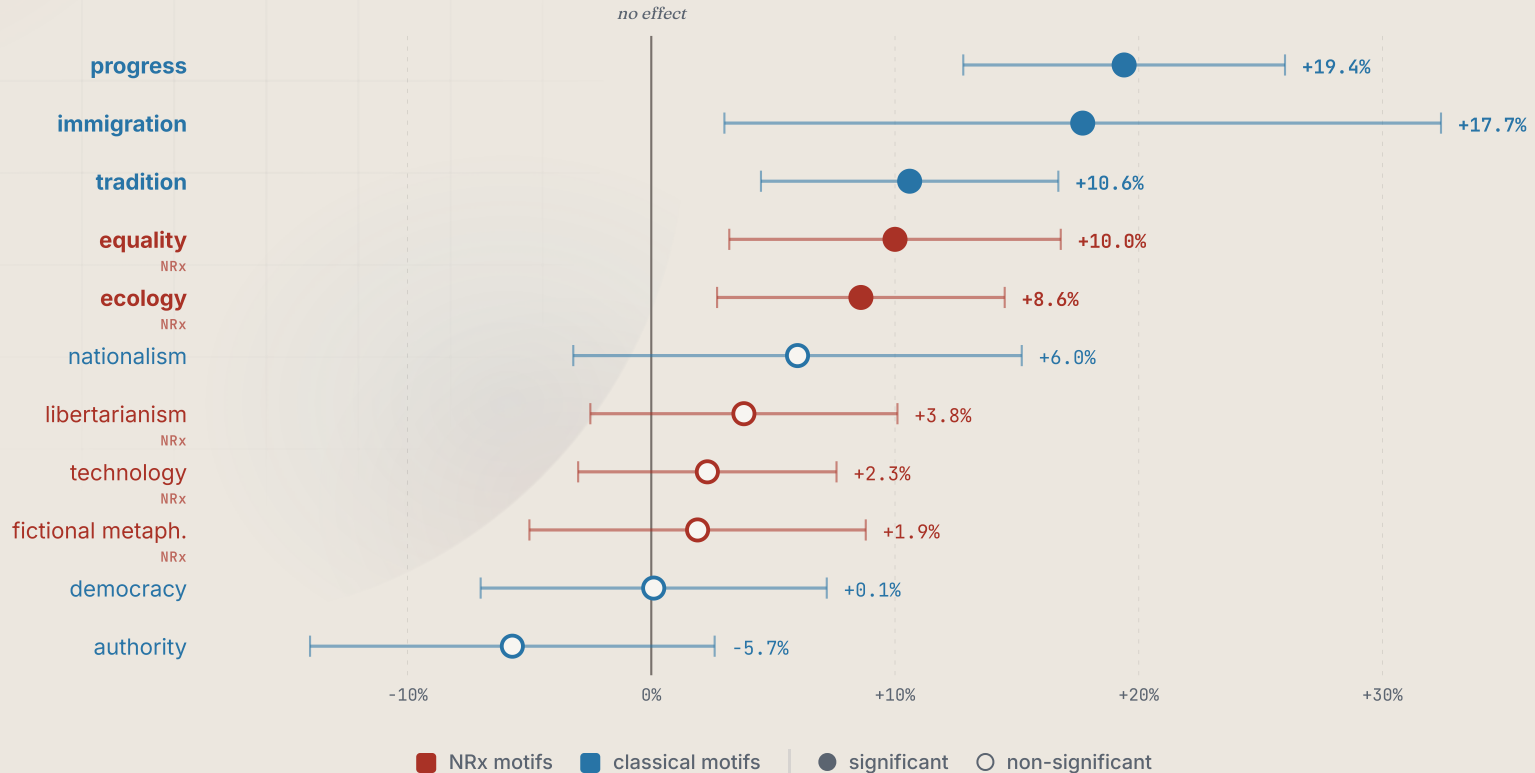


METHOD NOTE

Both curves rise. The audience-weighted curve plateaus around 2022 (existing voices saturate), but the channel-balanced curve keeps climbing: the drift is being driven by speakers themselves, not just by which channels grow.

Which themes politicize the audience?

Effect of each motif on comment politicization, +1 SD shift, PPML with channel fixed effects, 95% CI



Limitations and next steps

1 TOP-DOWN, YOUNPOL TODAY

Channels are identified a priori through expert assessment and audience metrics. This gives analytic clarity and reproducibility, but risks confirmation bias and may miss ambiguous or emergent actors.

2 BOTTOM-UP BRIDGE

The NER plus ideology pipeline now lets YouPol surface the actors that tracked creators cite, endorse, or engage with, and expand the corpus from inside the transcripts themselves.

INDEPENDENT TECHNICAL LIMITATIONS

a.

transcription

errors on overlapping speech, heavy accents, degraded audio

b.

silver labels

LLM-in-the-loop is not gold-standard; iteration mitigates but does not eliminate noise

c.

platform dependence

initial collection still relies on platform availability

d.

linguistic scope

francophone today, anglophone expansion in preparation

→ NEXT

Anglophone expansion, TikTok parity with the YouTube subset, and an open question: does political density predict subsequent platform removal?

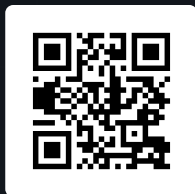
Thank you!

Questions and discussion

Tristan Boursier · Antoine Lemor, Ph.D.

Sciences Po Paris · UQAM · Université de Sherbrooke

CIRST · RFICS · CAPP



you-pol.com

PROJECT



data.you-pol.com

APSA2026