

EUROPEAN POLITICAL SCIENCE SOCIETY ANNUAL CONFERENCE, BELFAST, JUNE 18-20, 2026

LLM Tool

Improving transparency and validity in computational text analysis

A local hybrid pipeline: LLM annotation · transformer classifiers · validation logs

Antoine Claude Lemor¹, Shannon Dinan², Jérémy Gilbert³

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



LLM Tool · GitHub

Five minutes, three claims

1

Context

LLMs help, but threaten reproducibility and validity

2

Method

LLM Tool as a local, logged hybrid workflow

3

Validation

human consensus, model comparison, speed and traceability

PART I

Context and literature

LLMs make annotation easier, but validity and reproducibility remain the methodological problem.

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?

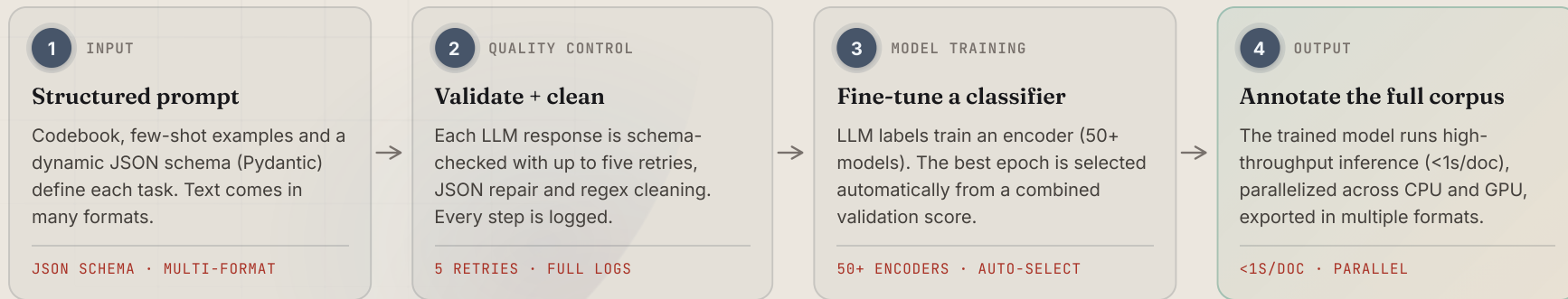
PART II

Method and tool

LLM Tool turns prompting, cleaning, training and inference into one logged local workflow.

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

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



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.

What makes the tool transparent: modes plus persistent records

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.

RECORDED INPUTS

prompts, concept definitions, labels, examples, JSON schema, provider and model version.

RECORDED EXECUTION

temperature, seeds, retries, raw outputs, cleaned outputs, validation failures, runtime and cost logs.

MINIMUM REPRODUCIBILITY

Even when inference is stochastic, the analytical choices are auditable and the run can be repeated under documented conditions.

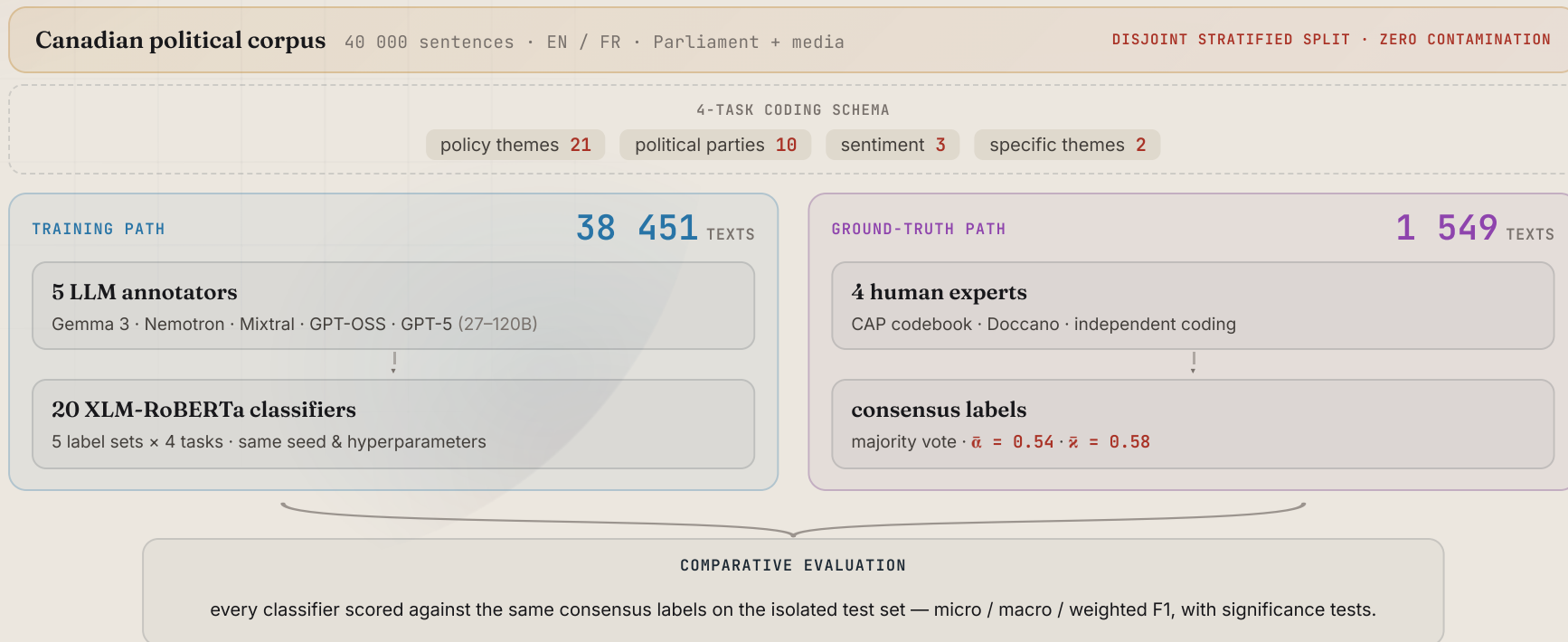
PART III

Validation results

The test is not whether LLMs sound plausible, but whether their labels survive human-grounded validation.

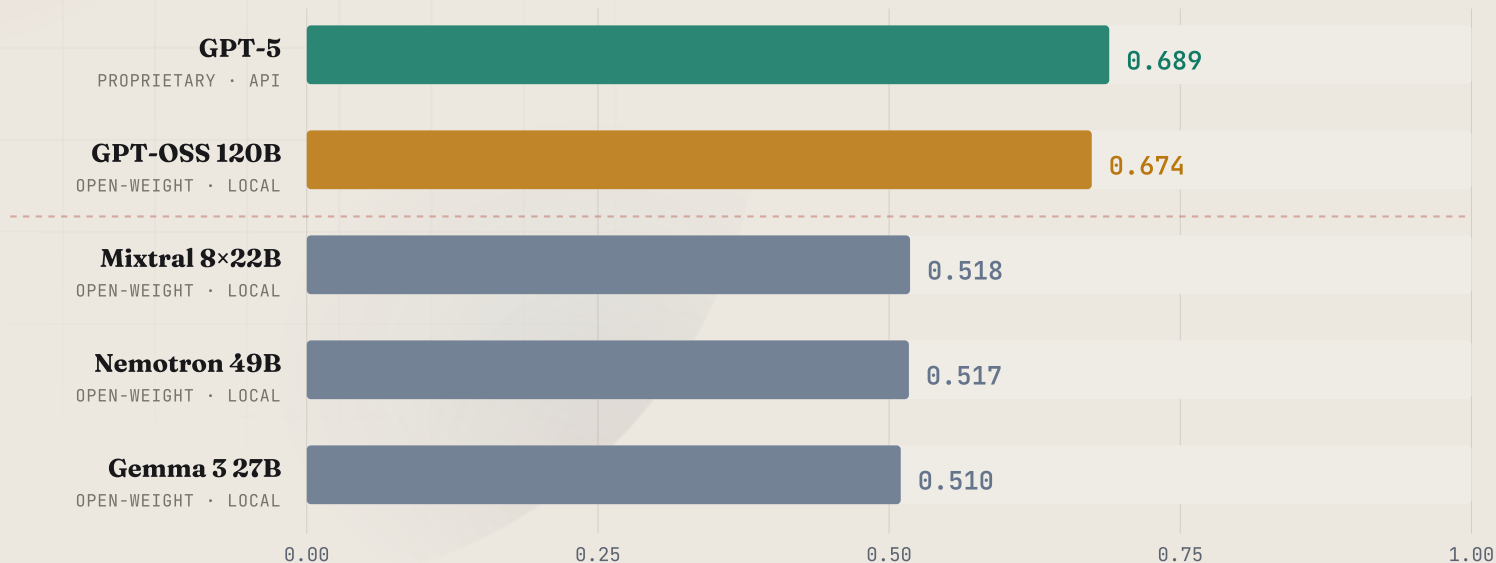
Validation design: isolate the human test set

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



Validation results: two tiers, one local competitor

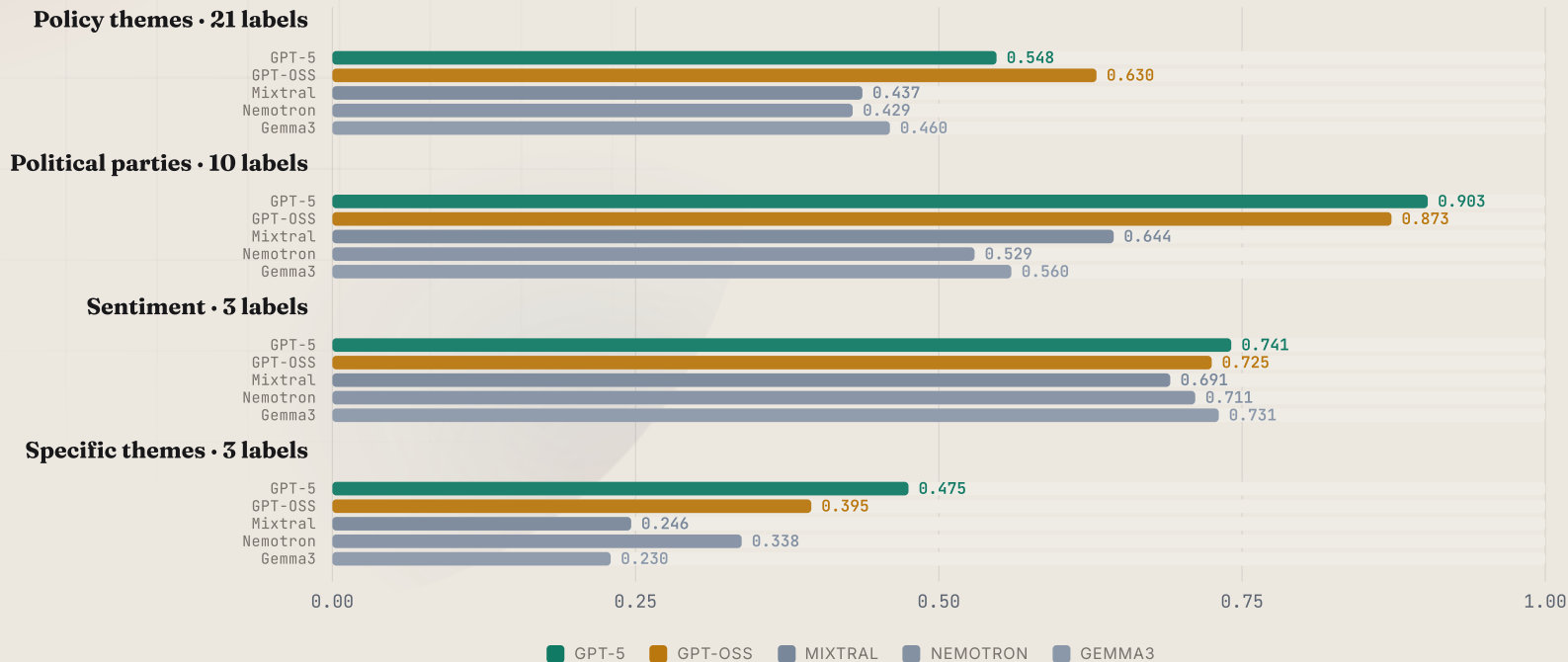
Mean Micro F1 across four annotation tasks, evaluated on the isolated human-consensus test set (N = 1,549)



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

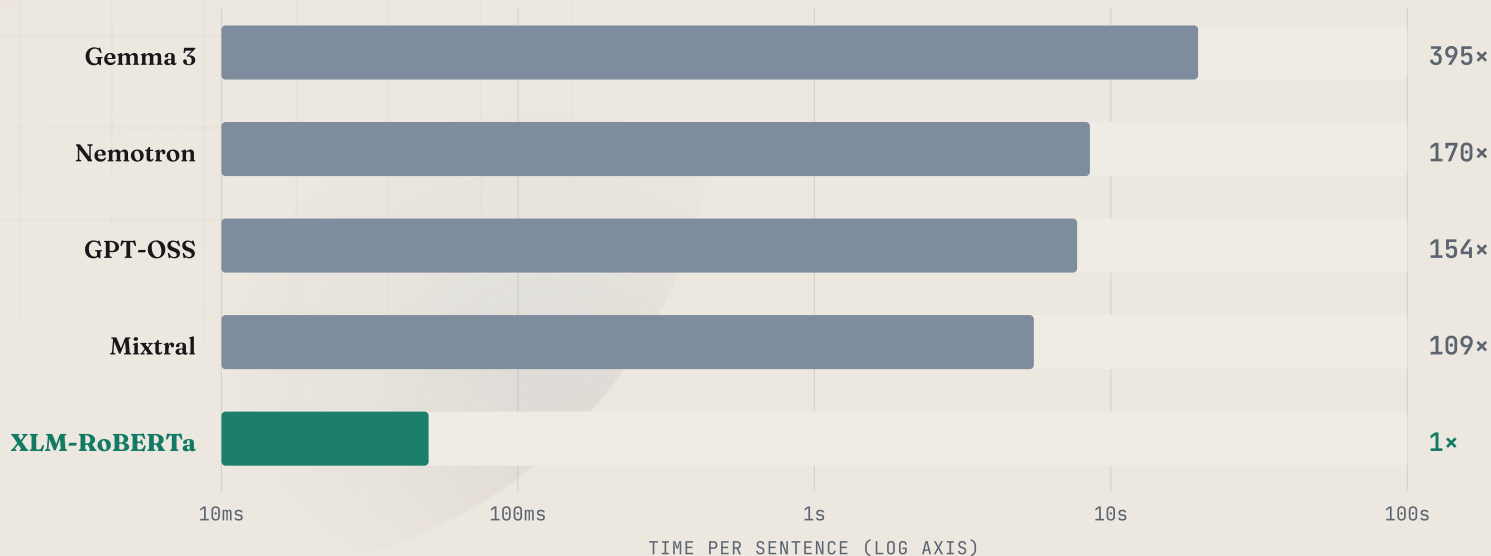
Micro F1 by annotation task and source model, evaluated on the same human-consensus test set



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.

LLM TOOL

Thank you

1

Validated

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

2

Transparent

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

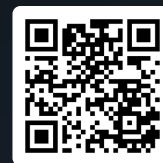
3

Open-source

The code, data and documentation are available so the workflow can be inspected, adapted and reused.

Antoine Claude Lemor · Shannon Dinan · Jérémy Gilbert

European Political Science Society · Belfast · 20 June 2026



LLM Tool · GitHub