Oxford Computational Political Science Group
Let Data Inspire.

UNIVERSITY OF OXFORD

DPIR
DEPARTMENT OF POLITICS & INTERNATIONAL RELATIONS

OCPSG Research Programme

Benchmarking LLMs and Fine-Tuned Models for Multilingual Policy Agenda Annotation

Bastián González-Bustamante, Tom Bellens, Christopher Klamm, and Marta Koch

ocpsg@politics.ox.ac.uk

Presentation at the OCPSG Research Cafe, March 27, 2026

Version of Record (VoR)

Project Overview & Motivation

Substantive Goal

Comparative political analysis increasingly requires scalable approaches to classify parliamentary interventions into policy topics across languages and institutional settings.

Complex Classification

Assigning one of 21 major policy topics to parliamentary speeches.

Methodological Goal

Recent LLM advances have raised new questions about robustness, comparability, interpretability, computational cost, and whether prompt-based systems can outperform supervised classifiers trained for this task.

The project treats multilingual topic annotation not merely as a prediction problem, but as a **data quality problem** in computational social science.

The Core Aim

> The aim is not only to identify the model that achieves the highest predictive score, but to construct a workflow capable of generating reliable, auditable, and scalable topic labels for multilingual parliamentary corpora.

Harmonised Data

Legislative text from standardised European corpora.

Consensus Framework

Competence-weighted silver-standard construction.

Human Validation

Progressive conversion of uncertain labels into a gold benchmark.

Data Sources & Corpus Architecture

ParlaMint 5.0

Our principal data source. Provides harmonised parliamentary corpora for a large set of European countries and regions, with a standardised structure supporting cross-national comparability. It includes both **original-language speeches** and **English machine translations**.

ParlaCAP

Extends ParlaMint with automatic policy-topic labels and sentiment estimates. Supplies an existing multilingual classification layer and serves as one benchmark baseline. Catch-all outputs such as "Mix" and "Other" must not be treated as equivalent to substantive residual classes.

ParlLawSpeech (Germany)

Used specifically for **German parliamentary speeches**. Source provenance is retained in metadata so that country-specific performance differences (reflecting corpus structure, preprocessing, or institutional language) can be interpreted carefully.

Three Baseline Classifiers

Stream 1: ParlaCAP

Existing topic-classification layer already linked to the parliamentary corpus. Provides multilingual labels directly comparable against other models.

Stream 2: CAP Babel Machine (MT)

Fine-tuned XLM-RoBERTa-large applied to English machine-translated speeches. Reduces complexity from language heterogeneity via a common channel.

Stream 3: CAP Babel Machine (original)

Same framework deployed on source-language speeches. Language-specific XLM-RoBERTa models used where available; multilingual variant otherwise.

For Germany, ParlaCAP and CAP Babel Machine in both multilingual and German variants were used.

Computational Implementation & Inference Efficiency

~100s

CPU Inference

Processing time per 100 observations on CPU in pilot runs

~1.3s

GPU Inference

Processing time per 100 observations after transition to CUDA-based GPU inference

This dramatic speed-up makes corpus-wide inference, repeated benchmarking, and multi-model agreement analysis operationally viable. Computational performance is treated as part of the evaluation architecture, not a secondary concern. **Carbon footprint data** has also been collected for incorporation into the comparative assessment.

Coverage across countries in the dataset.

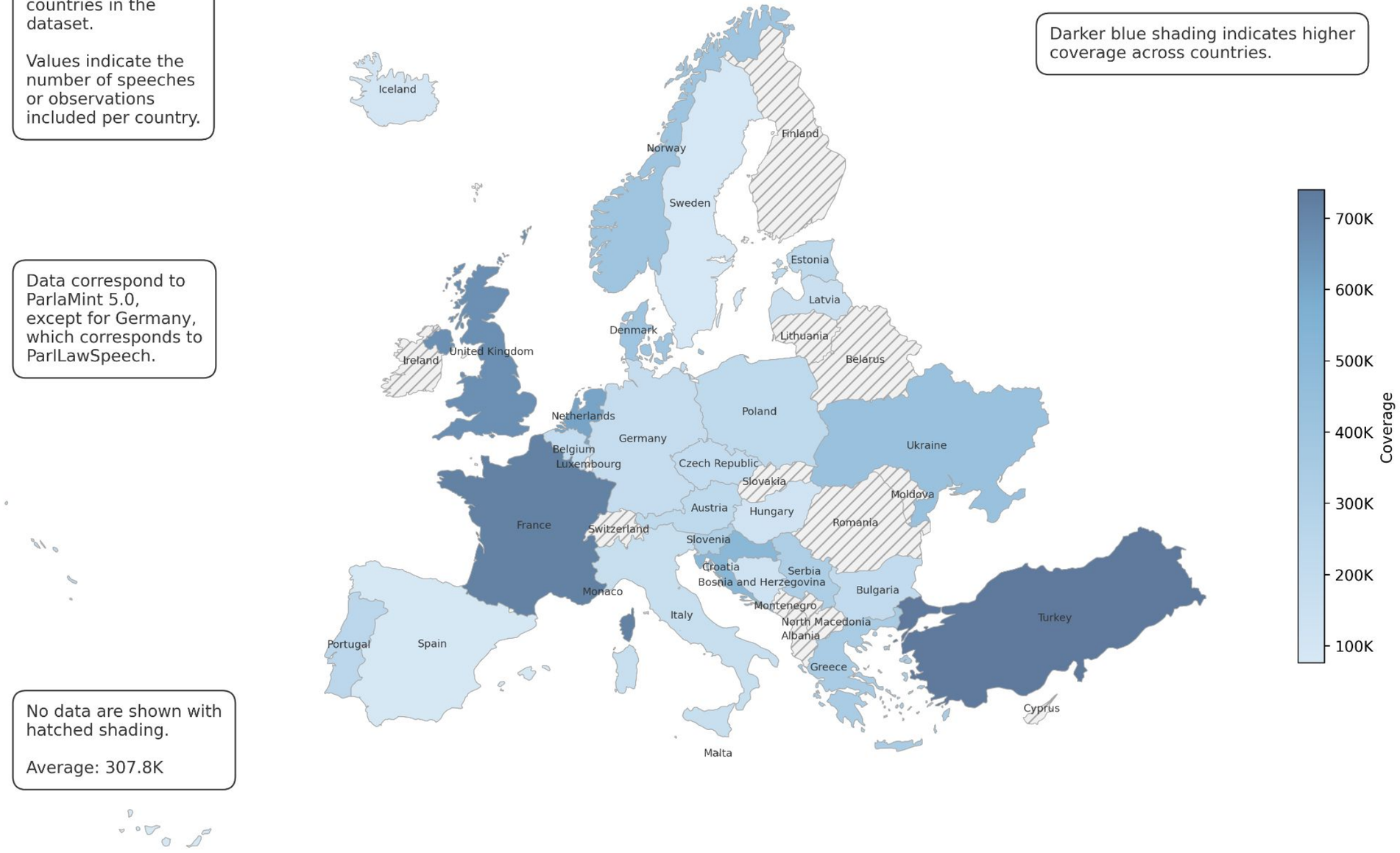
Values indicate the number of speeches or observations included per country.

Data correspond to ParlaMint 5.0, except for Germany, which corresponds to ParlLawSpeech.

No data are shown with hatched shading.

Average: 307.8K

Darker blue shading indicates higher coverage across countries.



Agreement between ParlaCap and Babel Machine across countries.

Values are reported as proportions.

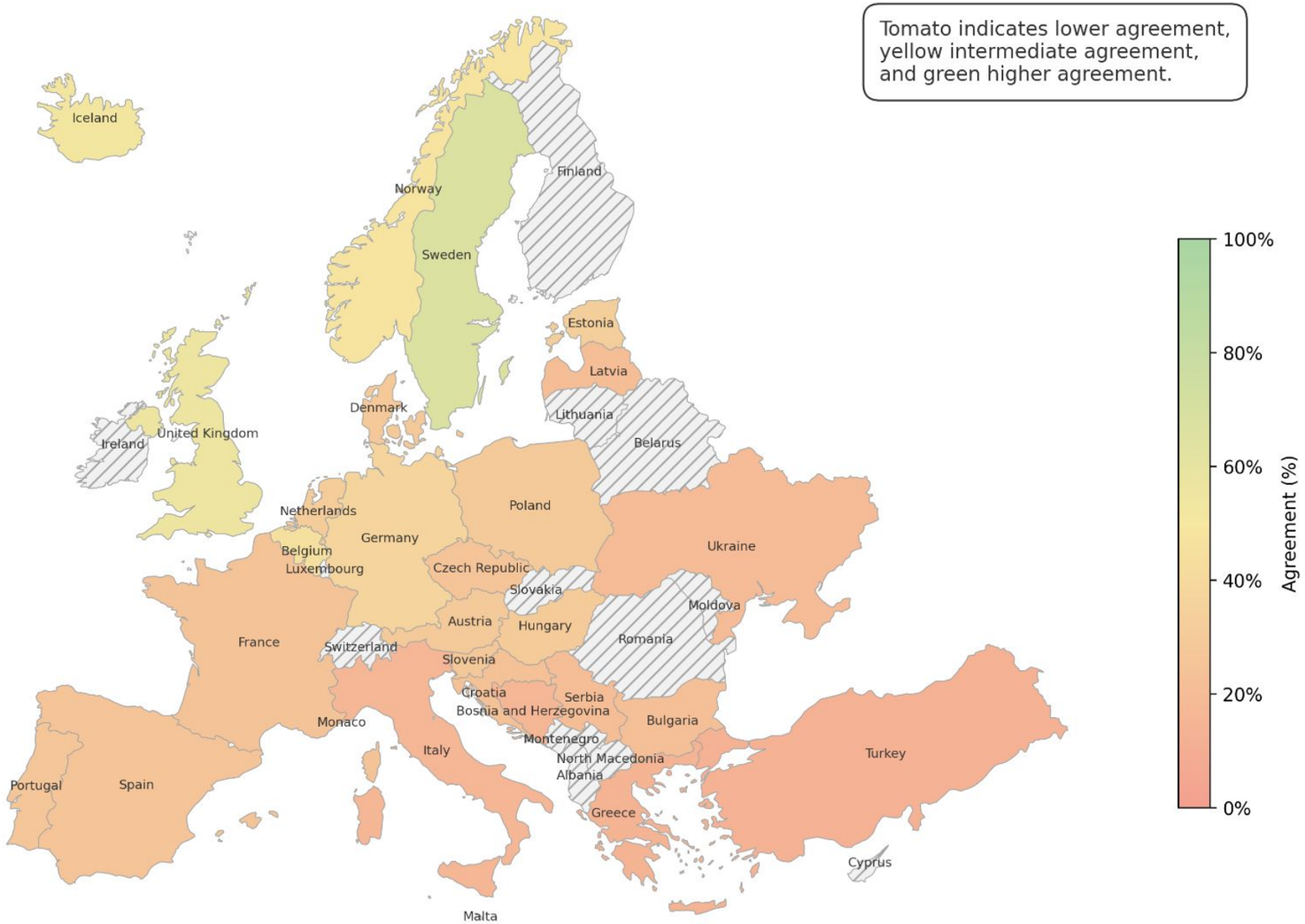
ParlaCap is an extension of ParlaMint that includes automatically annotated policy labels.

We ran Babel Machine (XLM-RoBERTa-large) on machine-translated speeches.

No data are shown with hatched shading.

Average: 30.5%

Tomato indicates lower agreement, yellow intermediate agreement, and green higher agreement.



Agreement across three classifiers by country.

Values are reported as proportions.

This measure extends the previous comparison by incorporating an additional classification using original-language speeches.

No data are shown with hatched shading.

Average: 23.8%

Tomato indicates lower agreement, yellow intermediate agreement, and green higher agreement.

CWCD: Competence-Weighted Consensus

A central methodological innovation. Rather than merging classifier outputs by naïve majority vote, the project aggregates them through **CWCD (beta package)**, developed from the logic of MACE, adapted for probability-aware model consensus.

Model Competence

Estimates which classifier is most reliable, on which labels, and to what degree using Variational Bayes EM.

Item Ambiguity

Identifies which observations are intrinsically ambiguous via consensus entropy.

Model Deviation

Measures how far each classifier departs from the inferred consensus using competence-penalised Jensen-Shannon divergence.

CWCD: Technical Design

From Hard Labels to Soft Consensus

CWCD extends MACE from hard-label aggregation to soft-label consensus estimation. Each speech becomes an item with three model-generated annotations. CWCD estimates a latent consensus distribution over labels, not a simple arithmetic reconciliation.

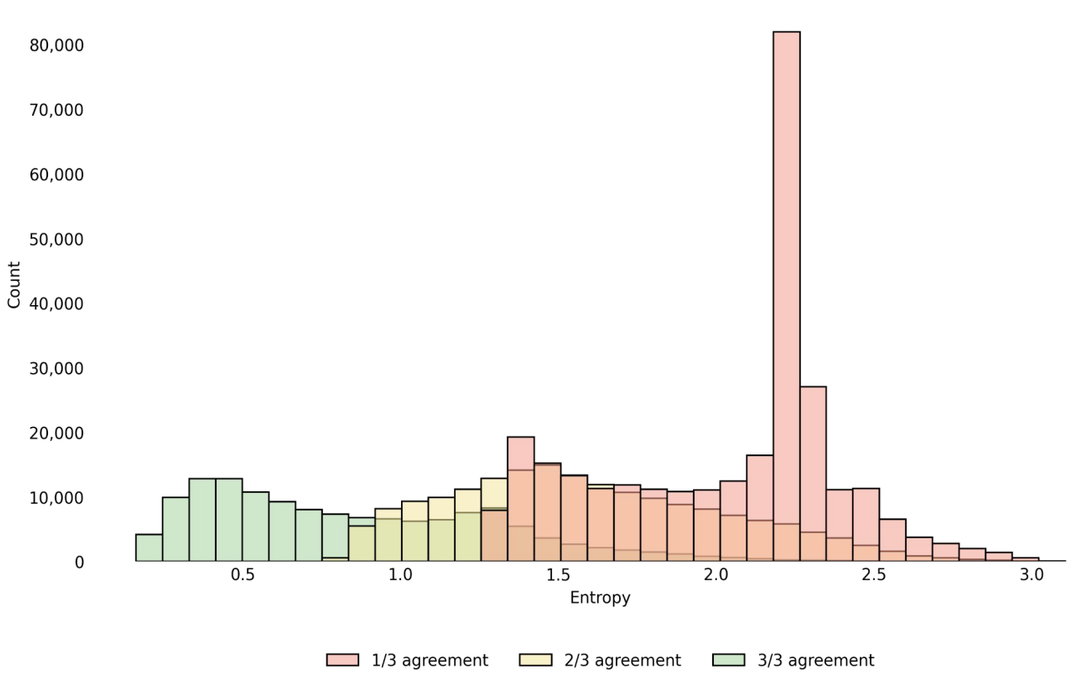
F1 Priors & Support Weighting

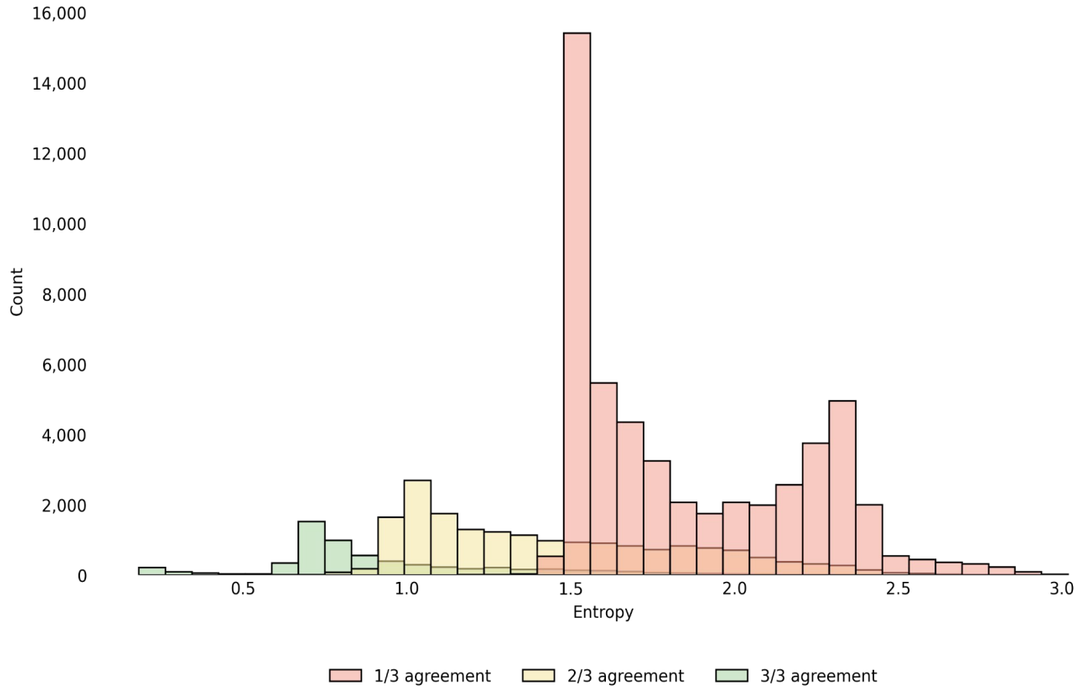
Per-label F1-scores and label support from the test set are incorporated as priors. This recognises that a classifier may be strong on some policy topics and weak on others. **Support-based shrinkage** ensures labels estimated on very small samples do not disproportionately influence the consensus.

- Agreement among three streams differing in model origin and language channel is more informative than repeated predictions from a single pipeline.

Entropy Distribution

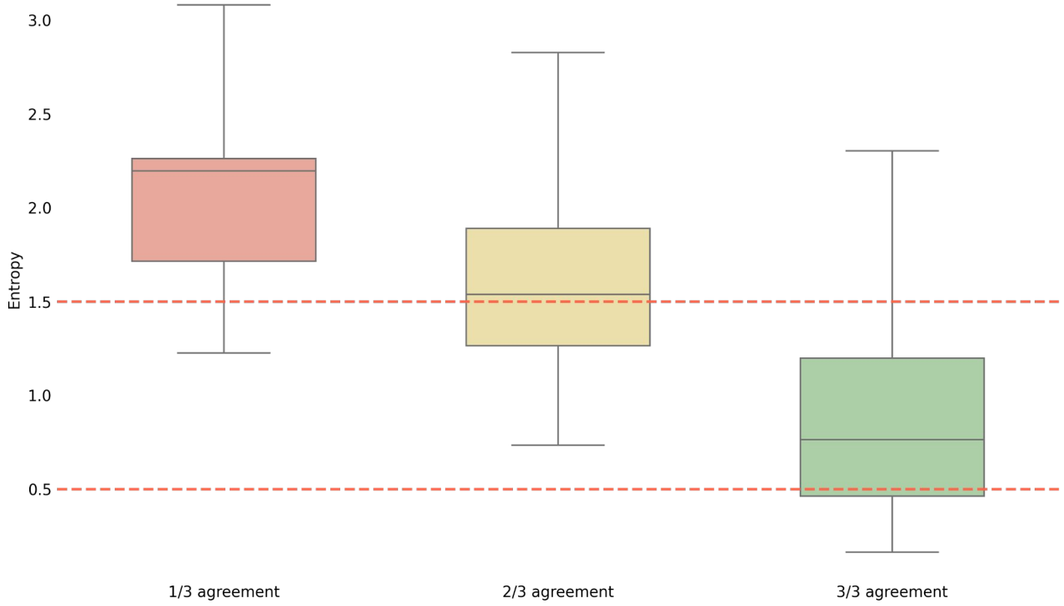
Netherlands and Spain

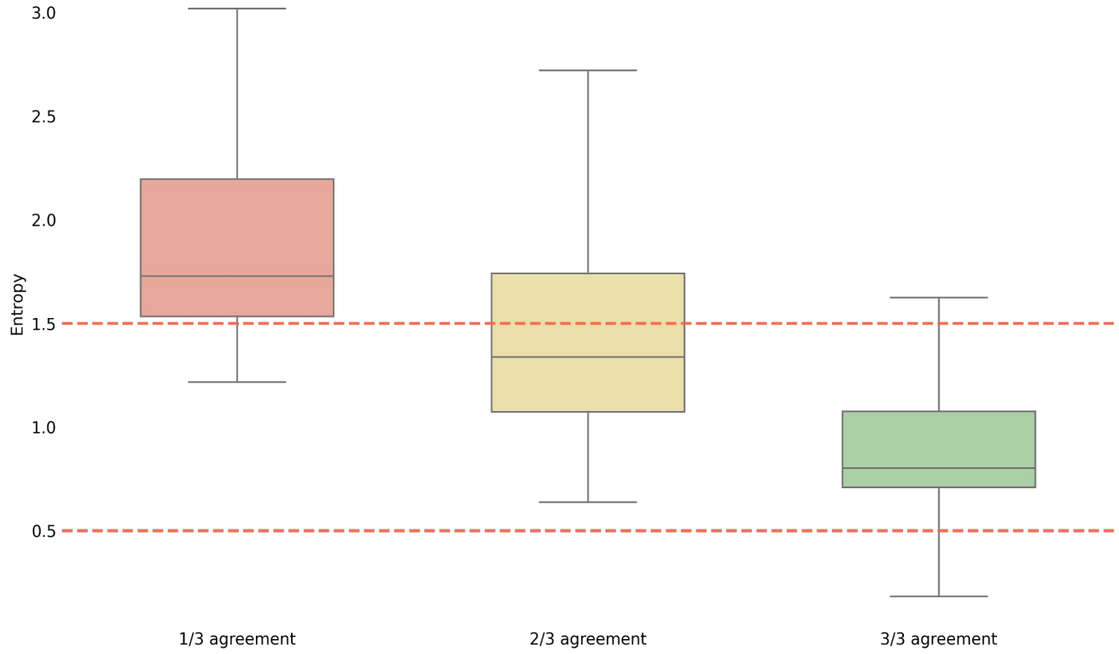


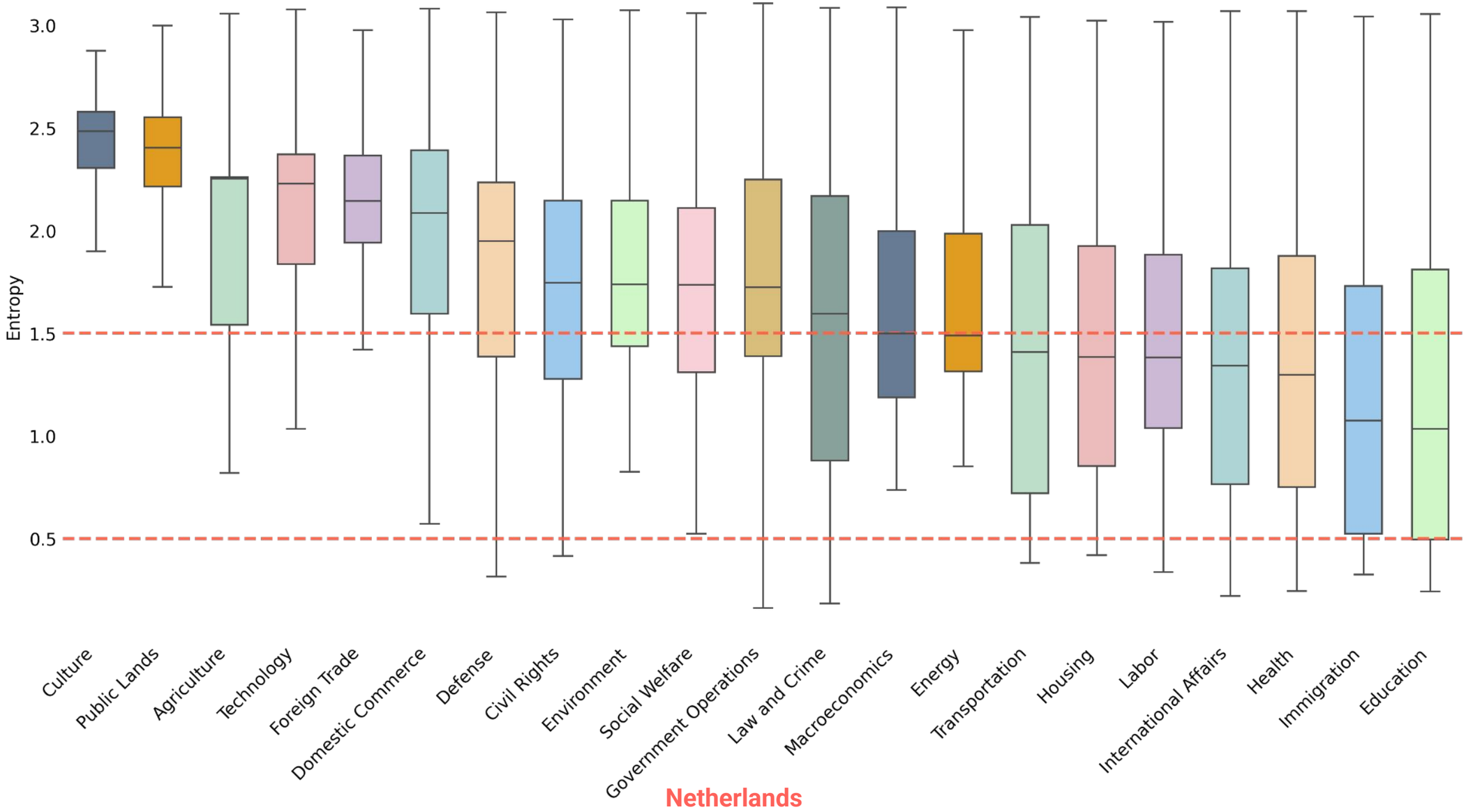


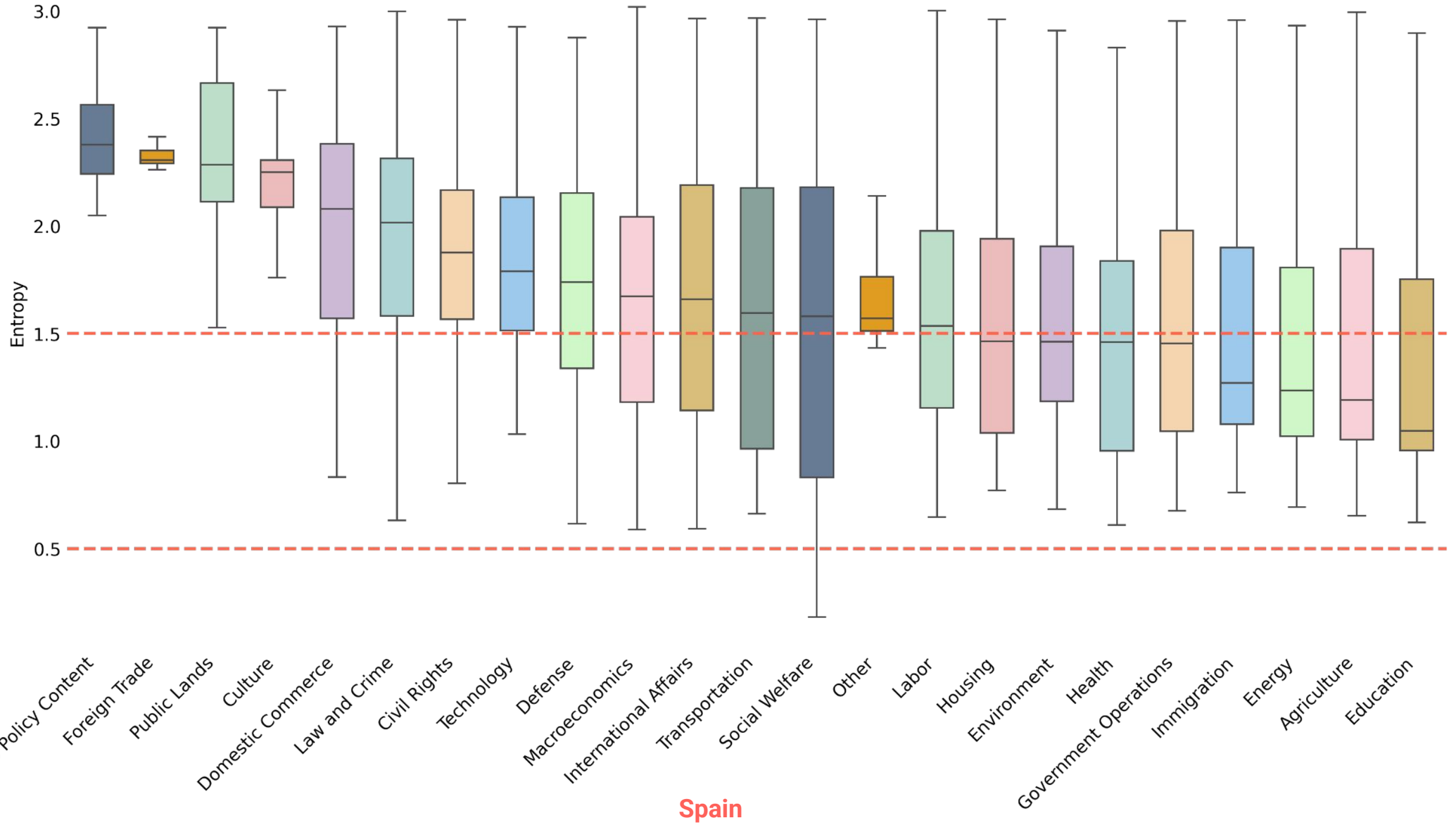
Entropy Boxplots by Agreement

Netherlands and Spain



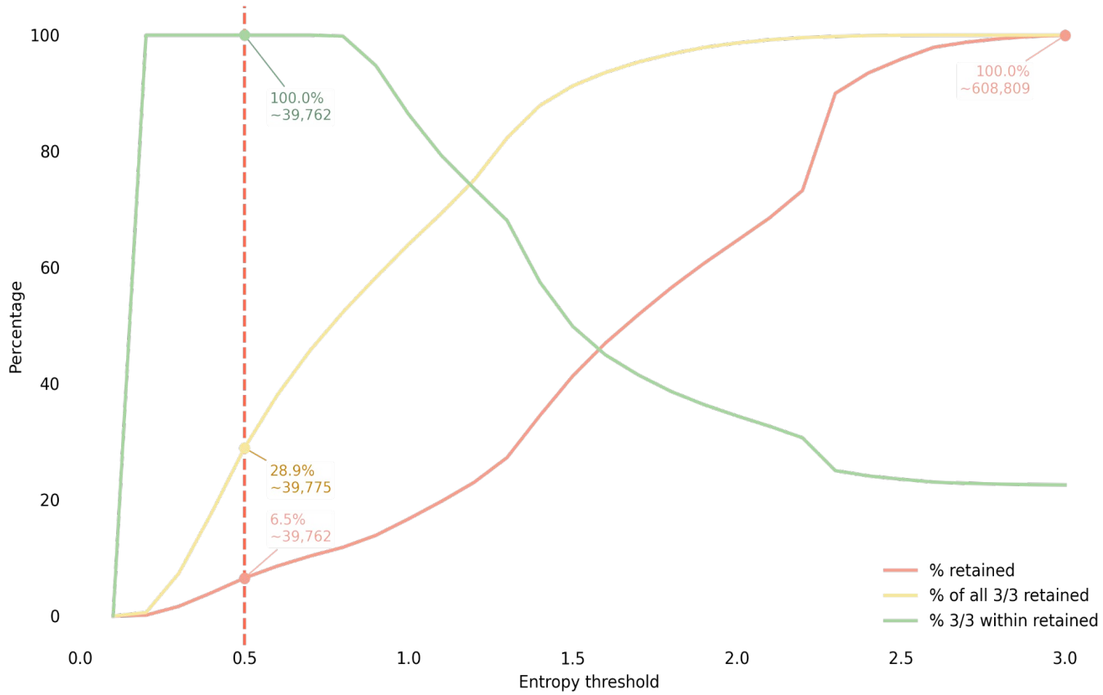


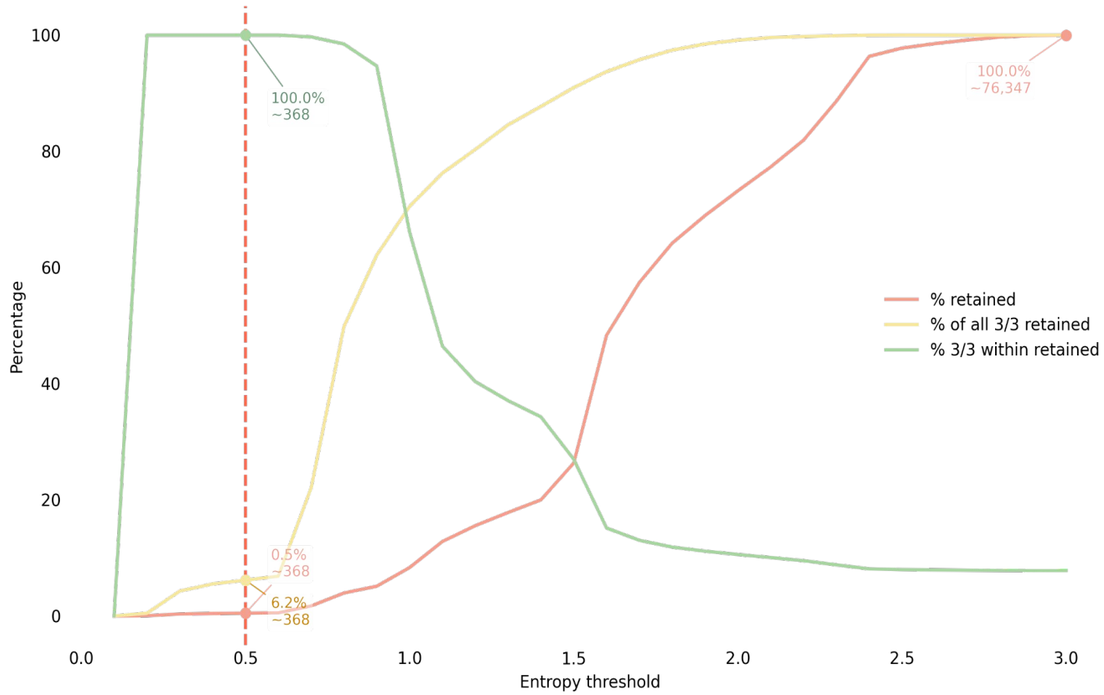

Entropy
3.0
2.5
2.0
1.5
1.0
0.5
Culture
Public Lands
Agriculture
Technology
Foreign Trade
Domestic Commerce
Defense
Civil Rights
Environment
Social Welfare
Government Operations
Law and Crime
Macroeconomics
Energy
Transportation
Housing
Labor
International Affairs
Health
Immigration
Education
Netherlands

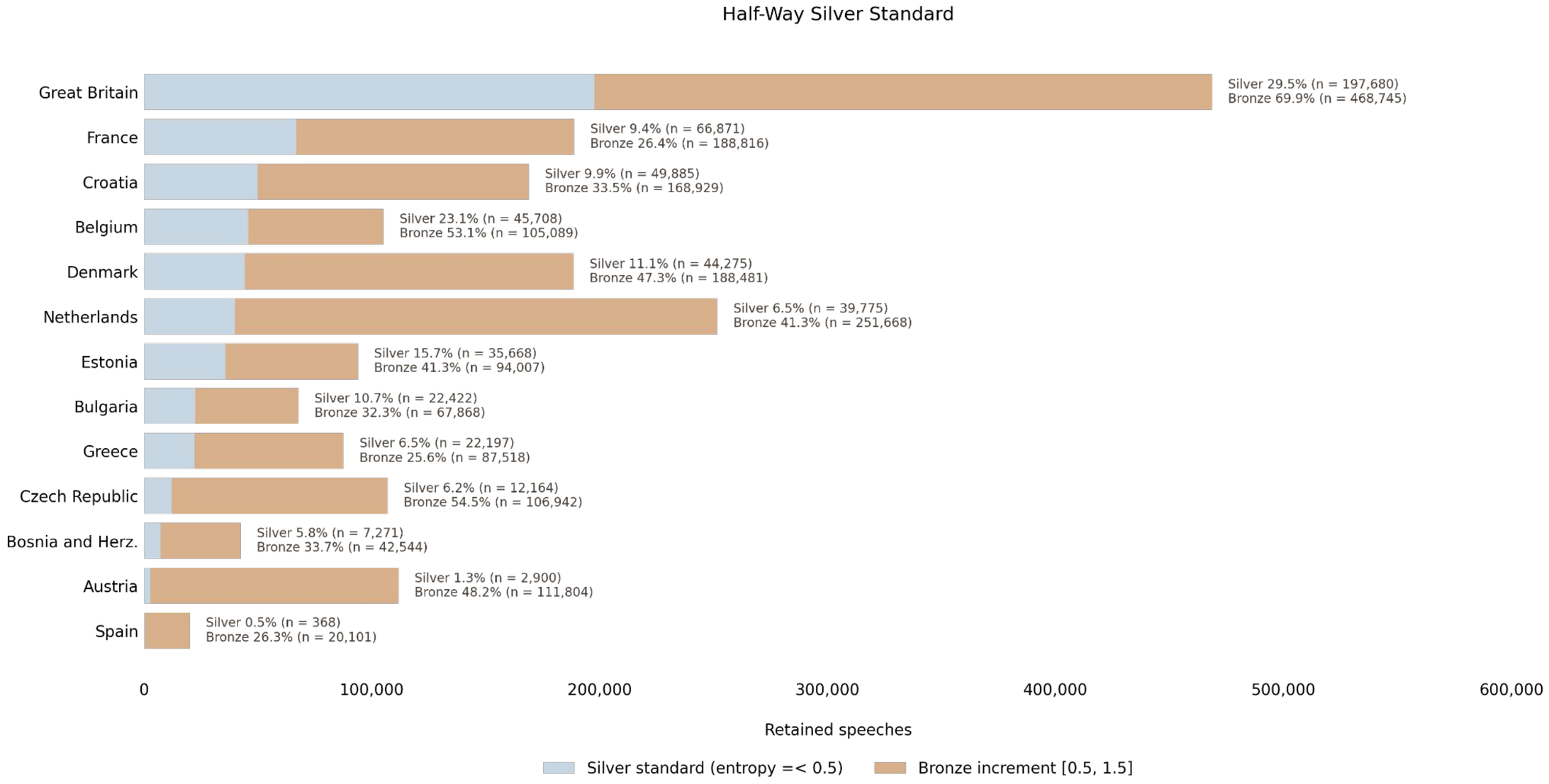

3.0
2.5
2.0
1.5
1.0
0.5
Entropy
No Policy Content
Foreign Trade
Public Lands
Culture
Domestic Commerce
Law and Crime
Civil Rights
Technology
Defense
Macroeconomics
International Affairs
Transportation
Social Welfare
Other
Labor
Housing
Environment
Health
Government Operations
Immigration
Energy
Agriculture
Education
Spain

Threshold Diagnostic

Netherlands and Spain







Sum of silver retained speeches = 547,184. Average silver retained percentage = 10.5%.
Additional bronze speeches recovered between thresholds 0.5 and 1.5 = 1,355,328.

Data Pipeline and Roadmap

Data Ingestion

Original and machine-translated parliamentary speeches from ParlaMint 5.0 and ParlLawSpeech

CWCD Consensus

Estimates competence-weighted consensus, incorporating model reliability, ambiguity, F1 priors, and label support

LLM Benchmarking

Zero-shot, few-shot, and reasoning-oriented prompting strategies are benchmarked on the silver test set

Iterative Evaluation & Fine-tuning

Silver and gold resources drive continuous evaluation and fine-tuning of models

Baseline Classification

Three BERT models (ParlaCAP, CAP Babel MT, CAP Babel Original) generate initial labels

Silver Standard Creation

Raw outputs are transformed into a silver standard, enriched with diagnostics, forming train/val/test splits

Human-in-the-Loop Validation

Silver observations are converted into a gold benchmark through expert human review

Benchmarking LLMs

Once the silver standard is defined, the project proceeds with direct LLM benchmarking under controlled settings.

Zero-Shot

No examples provided; tests raw model knowledge of policy-topic categories.

Few-Shot

Exemplars held outside the evaluation partition; tests in-context learning.

Reasoning-Oriented

Chain-of-thought or structured prompting strategies for complex cases.

Prompt variations are treated as part of the experimental design. Both **deterministic** and **stochastic decoding** regimes are benchmarked explicitly (temperature settings should shape reliability and reproducibility of outputs).

Current Progress & Next Steps

1

Completed

Parliamentary corpus architecture assembled; three baseline classifier streams implemented; GPU inference operational gains demonstrated; CWCD framework developed.

2

Immediate Next Step

Run CWCD systematically on the three baseline label sources to finalise the silver standard and construct train/validation/test splits accounting for imbalance and uncertainty.

3

In Parallel

Launch controlled LLM benchmarking experiments and begin uncertainty-driven human adjudication to progressively convert the silver benchmark into a gold evaluation set.

4

Final Output

A rigorous comparison of LLMs and fine-tuned Transformers, plus a reproducible multilingual workflow for policy agenda annotation in parliamentary speeches.