ParliaBench: An Evaluation and Benchmarking Framework for LLM-Generated Parliamentary Speech

Marios Koniaris, Argyro Tsipi Panayiotis Tsanakas
National Technical University of Athens
School of Electrical and Computer Engineering, Division of Computer Science
Iroon Polytechniou 9, Zographou Campus, 15780 Athens, Greece
Email: mkoniari@central.ntua.gr, argyrotsipi@gmail.com, panag@cs.ntua.gr

Abstract

Parliamentary speech generation presents specific challenges for large language models beyond standard text generation tasks. Unlike general text generation, parliamentary speeches require not only linguistic quality but also political authenticity and ideological consistency. Current language models lack specialized training for parliamentary contexts, and existing evaluation methods focus on standard NLP metrics rather than political authenticity. To address this, we present ParliaBench, a benchmark for parliamentary speech generation. We constructed a dataset of 448k speeches from UK Parliament to enable systematic model training. We introduce an evaluation framework combining computational metrics with LLM-as-a-judge assessments for measuring generation quality across three dimensions: linguistic quality, semantic coherence, and political authenticity. We propose two novel embedding-based metrics, Political Spectrum Alignment and Party Alignment, to quantify ideological positioning. We fine-tuned five large language models (LLMs), generated 28k speeches, and evaluated them using our framework, comparing baseline and fine-tuned models. Results show that fine-tuning produces statistically significant improvements across the majority of metrics and our novel metrics demonstrate strong discriminative power for political dimensions otherwise absent from conventional evaluation, while domain fine-tuning reveals a measurable trade-off between political authenticity and lexical diversity.

Keywords: Parliamentary Speech Generation,LLM Evaluation,Political Authenticity,Benchmark Evaluation

1. Introduction

Democracy thrives through debate. Democratic parliaments are open forums where elected representatives engage in arguments over policy (Back et al., 2021). These debates provide unique insights into political reasoning and ideological positioning. Researchers in political science and computational linguistics increasingly seek to understand and model parliamentary debates. Generating such speech reliably would support a range of research tasks, such as studying political bias, building training data for ideology classifiers, and modelling parliamentary dynamics. Each depends on evaluation methods that do not yet exist.

Yet generating authentic parliamentary speech presents significant challenges that extend well beyond typical text generation tasks. The complexity becomes apparent when considering specific examples. A Labour MP discussing taxation policy must sound distinctly different from a Conservative counterpart not just in policy position, but most significantly, in their fundamental approach to governance. Large language models have shown promise across political applications, from sentiment analysis (Bestvater and Monroe, 2023) and election forecasting (Li et al., 2024b) to synthetic survey data generation (Argyle et al., 2023; Bisbee et al., 2024). Authentic parliamentary speech generation, however, remains challenging.

Evaluation approaches for generated political text fall short for these specialized requirements.

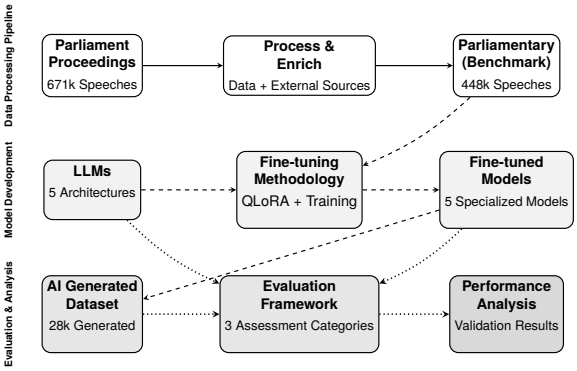


Figure 1: Methodology Overview: Data processing (top row) creates Dataset A from UK Parliament proceedings. Model development (middle row) fine-tunes base architectures using Dataset A. Evaluation pipeline (bottom row) applies the assessment framework to generated speeches (Dataset B) for performance analysis.

Although evaluation methods for text generation tasks have evolved from simple overlap metrics (Papineni et al., 2002; Lin, 2004) toward more sophisticated embedding-based approaches (Zhang et al., 2020; Zhao et al., 2019), they still focus on surface-level similarity rather than political authenticity.

Similarly, domain-specific benchmarks have emerged for legislative summarization (Kornilova and Eidelman, 2019), opinion alignment (Santurkar

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4797–4818
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

et al., 2023), and other specialized fields (Xie et al., 2024; Li et al., 2024a). Parliamentary speech generation, however, lacks the evaluation framework needed to assess ideological consistency and parliamentary conventions simultaneously.

Contributions. We address these limitations by establishing a benchmark resource designed specifically for evaluating parliamentary speech generation. First, we develop a curated dataset containing 448k speeches from UK Parliament (Section 3). Second, we introduce a multi-dimensional evaluation framework that assesses both linguistic quality and political authenticity (Section 4). Third, we fine-tune five large language models and generate 28k parliamentary speeches to establish baseline performance (Section 5). Finally, we demonstrate the framework's effectiveness through systematic evaluation (Section 6). Figure 1 provides an overview of the complete ParliaBench framework and experimental methodology.

2. Related Work

Recent advances in large language models (LLMs) have enabled a wide range of applications across political domains. (Li et al., 2024b) outlines several applications of LLM in political contexts, covering predictive, generative, and simulation-based approaches. The use of LLMs as substitutes for human experts in annotating political texts across multiple languages is explored in (Heseltine and von Hohenberg, 2024), while (Gunes and Florczak, 2023) employ LLMs for classifying U.S. Congressional bills. (Argyle et al., 2023) investigate LLMs as proxies for specific human subpopulations in social science research and (Bisbee et al., 2024) raise concerns about the quality, reliability, and reproducibility of synthetic survey data generated by LLMs. Agent-based LLMs are utilized as coalition negotiators (Moghimifar et al., 2024) and as U.S. senators simulating legislative processes (Baker and Azher, 2024). However, these approaches predominantly emphasize analytical and simulation capabilities rather than authentic speech generation quality.

Traditional text generation evaluation has evolved from reference-based metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) toward embedding-based approaches that better capture semantic similarity. BERTScore (Zhang et al., 2020) uses contextualized embeddings to compute token-level similarity, while MoverScore (Zhao et al., 2019) measures semantic transportation cost using Earth Mover's Distance. For reference-free evaluation, Zhu and Bhat (Zhu and Bhat, 2020) propose GRUEN, assessing grammaticality and semantic coherence. Domain-specific datasets like BillSum (Kornilova and Eidelman, 2019) for legislative summarization and OpinionQA (Santurkar et al., 2023) for opinion alignment provide targeted evaluation resources, though gaps remain in generative parliamentary speech assessment.

The emergence of LLM-as-a-Judge evaluation (Zheng et al., 2023) offers scalable alternatives for nuanced assessment, achieving over 80% agreement with human evaluators in complex judgment tasks. (Liu and Sun, 2023) further validate this approach, demonstrating GPT-4's high alignment with human thematic coding in political analysis. Likewise, (Bojić et al., 2025) show that LLMs achieve higher inter-rater reliability than human annotators for both sentiment and political leaning assessment, supporting the validity of automated political authenticity evaluation. This methodology has been successfully applied across diverse contexts, from general LLM benchmarking through competitive debates (Moniri et al., 2025) to long-context reasoning evaluation in parliamentary debates (Tiwari et al., 2025). These approaches demonstrate the viability of automated evaluation for argumentative and political content, though models exhibit documented biases toward Western, educated populations (Durmus et al., 2023) and systematic preferences in political simulations (Qi et al., 2024).

Regarding parliamentary speech data, structured corpora like ParlaMint (Erjavec et al., 2025) provide multi-lingual parliamentary proceedings. Embedding-based approaches for political analysis, introduced in (Rheault and Cochrane, 2020), demonstrate that embeddings can capture ideological positioning in parliamentary text. While stance detection research has largely focused on social media contexts (ALDayel and Magdy, 2021), ideology classification from legislative speech (Biessmann et al., 2016; Lauderdale and Herzog, 2016) is more directly relevant to parliamentary text and has likewise treated political positioning as a classification problem. We introduce metrics that utilize embeddings for generative evaluation, measuring how faithfully LLM-produced speech reproduces ideological positioning relative to ground-truth party centroids, a direction unaddressed by prior classification-focused work.

Domain-specific evaluation frameworks have emerged across professional fields, including FinBen (Xie et al., 2024) and LexEval (Li et al., 2024a). Political science applications have developed specialized benchmarks for election prediction and legislative analysis, yet these focus primarily on classification and analysis tasks. Parliamentary speech generation has attracted recent computational interest, with work exploring European Parliament consensus building (Zhang et al., 2025) and political impersonation authenticity (Herbold et al., 2024), but evaluation frameworks remain under-

developed. Existing approaches focus on narrow aspects like style mimicry rather than systematic quality assessment across linguistic and political authenticity dimensions that parliamentary speech generation requires.

Our work addresses these gaps by establishing a benchmark resource specifically designed for parliamentary speech generation.

3. ParliaBench Dataset

3.1. Construction & Processing Pipeline

We constructed our dataset from the UK subset of the ParlaMint corpus (Erjavec et al., 2025), using the raw version to retain preprocessing control and apply a four-step processing pipeline:

Step 1: Data Ingestion & XML Parsing. Parsed XML files to isolate individual speeches while preserving hierarchical organization.

Step 2: Metadata Extraction & Temporal Alignment. Enriched speeches with speaker identity, political affiliation, chamber designation, and session dates. As parliamentary speakers frequently change affiliations and roles during their careers, we employed temporal alignment by cross-referencing speech dates with affiliation histories from corpus metadata.

Step 3: Content Processing & Filtering. Distinguished substantive content from procedural elements, separating parliamentary prompts from speech responses. Filtered procedural noise and non-substantive speeches. For political affiliations, we applied 1000-speech minimum threshold, reducing from 28 original affiliations to 11 to ensure stable model training.

Step 4: Thematic Classification. While ParlaMint uses CAP classification, we selected EuroVoc (Publications Office of the European Union, 2025) as the standard classification system for European parliamentary systems. For policy domains with clear semantic correspondence between CAP and EuroVoc taxonomies, we applied direct mapping rules. For semantically complex or ambiguous categories, we employed the methodology provided by (Bocchi et al., 2024). We argue that this approach is particularly well-suited for our dataset because it was specifically designed for legal and governmental texts. For speeches yielding multiple concepts, we selected the highest individual concept score. (Appendix 9.2 documents the hybrid classification strategy)

3.2. Statistics

The final dataset contains 447,778 speeches from 1,901 unique speakers across 11 political affiliations, with major parties (Conservative: 263,513; Labour: 108,831) dominating representation. The dataset contains approximately 99.94 million words, with speeches averaging 223 words (median: 99 words). This distribution reflects natural variation in parliamentary speeches, from brief procedural statements to extended policy expositions. Temporal coverage captures significant political events including Brexit debates, and the COVID-19 pandemic response, ensuring exposure to diverse political contexts and rhetorical situations. Note: "Bishops", "Crossbench", and "Non-Affiliated" are not political parties in the traditional sense but formal affiliations in Parliament.

Figure 2 illustrates key dataset characteristics. Panel (a) reveals a highly right-skewed distribution of speech lengths, indicating that the dataset is dominated by relatively short speeches while containing a smaller number of substantially longer ones. Panel (b) presents topic distribution showing that speeches most frequently address International Relations, Law, and Politics, which together account for over 45% of the corpus. Panel (c) demonstrates ideological balance across the political spectrum. While most parties by count are in the centre-left to left spectrum, the centre-right to right spectrum produces a larger amount of speeches by 110,000. Panel (d) displays temporal patterns and institutional differences. The House of Commons consistently produces 3-4x more speeches than the House of Lords across 2015-2022.

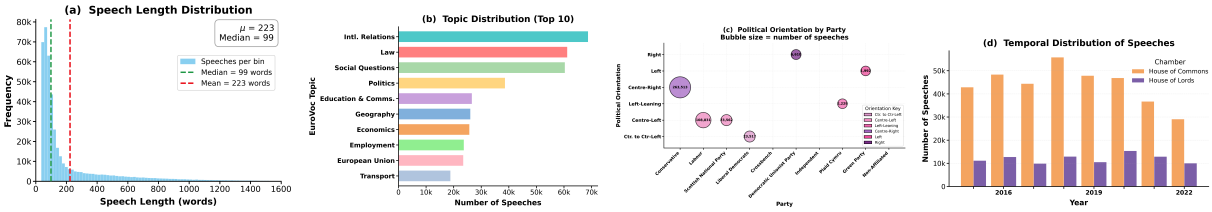
3.3. Format and Availability

The ParliaBench Dataset is distributed as two JSONL datasets:

Training Dataset. An example entry illustrates the dataset structure:

```
{
  "Party": "Labour Party", # Political affiliation
  "Topic": "Economy",      # Thematic classification
  "Political Orientation": "Left", # Party Pol. Or.
  "Section": "Economic Affairs",   # Parliamentary
     debate section
  "House": "House of Lords",   # Chamber of
     Parliament
  "Speech": "Thank you, Mr. Speaker. I rise to
     address...", # Complete transcript
  "Prompts": "What measures will your party
     take..." # Associated prompts
}
```

Generated Dataset. We release 27,560 speeches (model outputs) produced during evaluation (Section 5). The generated dataset follows the same format with additional fields: *Model* (architecture identifier), *Type* (baseline/fine-tuned), *Generated Speech* (model output), and *Evaluation Scores* (computed metrics). To ensure a fair comparison, we used the same input prompts across all models and model types, allowing evaluation on identical inputs. Both datasets and fine-tuned models are publicly available (see Section 7).



(a) (b) (c) (d)

Figure 2: Dataset statistics: (a) Speech length distribution, (b) Topic distribution, (c) Political orientation distribution, (d) Temporal distribution.

4. Evaluation Framework

Parliamentary speech evaluation requires assessment across multiple levels that generic benchmarks cannot capture. Our framework operates on two levels: (i) **speech evaluation metrics** measuring generation quality across three dimensions, and (ii) **consistency evaluation metrics** measuring performance reliability across political contexts. This dual-track approach also combines computational metrics with LLM-judge evaluation.

Our framework introduces two novel computational metrics (Political Spectrum Alignment and Party Alignment, Section 4.1.1) alongside established metrics from both computational and LLM-judge traditions.

Computational metrics provide deterministic assessment through established NLP measures, while LLM-judge metrics capture nuanced qualities requiring contextual understanding. For LLM-judge evaluation, we adapt the methodology from (Zheng et al., 2023). We employed Flow-Judge-v0.1 as our LLM judge, an LLM specialized in system evaluation tasks. It inherits it's architecture from Phi-3.5-mini instruct (Abdin et al., 2024), thus ensuring complete architectural and training data independence from the evaluated models. Our judge rates each speech on a 1-10 scale with written explanations across six parliamentary specific dimensions. We acknowledge this introduces bias through single model judgment. We employ consistent prompt formatting and evaluate speeches sampled across diverse political contexts. Specific prompts for each dimension are provided in Appendix 10.

Table 1 shows our evaluation structure. We assess three quality dimensions: (i) linguistic quality, (ii) semantic coherence, and (iii) political authenticity using both computational and LLM-judge metrics.

4.1. Speech Evaluation Metrics

1. Linguistic Quality

(a) **Computational Metrics**

i. **Perplexity** (PPL) (Jelinek et al., 1977) measures text naturalness. We compute perplexity using GPT-2 as a fixed reference model across all generated speeches, ensuring cross-model comparability. Lower scores indicate more natural-sounding text according to GPT-2's language distribution. (↓ better)

ii. **Distinct-n** (Dist-N) (Li et al., 2016) measures lexical diversity through unique bigram proportions (↑ better), capturing variety in phrase transitions without repetition.

iii. **Self-BLEU** (Zhu et al., 2018) measures intra-model diversity by computing BLEU similarity among variants generated under the same prompt. Lower scores indicate greater output variety. (↓ better)

(b) **LLM-Judge Metrics**

i. **Coherence** (J_Coh) evaluates logical argument flow and structural organization (↑ better).

ii. **Conciseness** (J_Conc) assesses communication efficiency, drawing from debate evaluation criteria (Moniri et al., 2025) (↑ better).

2. Semantic Coherence

(a) **Computational Metrics**

i. **GRUEN** (Zhu and Bhat, 2020) evaluates linguistic quality through grammatical correctness and semantic coherence (↑ better), aggregating four quality dimensions: grammaticality, non-redundancy, focus, and structural coherence.

ii. **BERTScore** (Zhang et al., 2020) measures semantic similarity using RoBERTa-large embeddings (↑ better), reporting F1-score to balance precision and recall.

iii. **MoverScore** (Zhao et al., 2019) quantifies semantic distance via Earth Mover's Distance (↑ better).

Note: For both BERTScore and MoverScore, generated speeches are compared against the top-5 human speeches from the training set matching the same context.

(b) **LLM-Judge Metrics**

i. **Relevance** (J_Rel) evaluates whether responses address given prompts (↑ better).

3. Political Authenticity

(a) **Computational Metrics**

i. We employ two novel embedding-based metrics for political authenticity assessment: Political Spectrum Alignment (PSA) and Party Alignment

Evaluation Level	Assessment Category	Computational Metrics	LLM-Judge Metrics
Speech	Linguistic Quality	PPL, Dist-N, Self-BLEU	J_Coh, J_Conc
	Semantic Coherence	GRUEN, BERTScore, MoverScore	J_Rel
	Political Authenticity	PSA, Party Align	J_Auth, J_PolApp, J_Qual
Consistency	Cross-Context Stability	Applied to all speech evaluation metrics	

Table 1: Multi-Dimensional Evaluation Framework.

(Party Align). Detailed calculation methodology provided in Section 4.1.1.

(b) **LLM-Judge Metrics**

i. **Authenticity** (J_Auth) assesses whether content reflects genuine political speech (↑ better).

ii. **Political Appropriateness** (J_PolApp) evaluates whether tone is suitable for political speech (↑ better).

iii. **Overall Quality** (J_Qual) assesses sophistication, persuasiveness, and communicative effectiveness (↑ better).

4.1.1. Novel Political Authenticity Metrics

Political Spectrum Alignment (PSA) evaluates ideological positioning on the left-right spectrum. We adapt semantic embedding approaches (Rheault and Cochrane, 2020) for LLM-generated speech evaluation, drawing on the Left-Right (RILE) scale methodology (Volkens et al., 2013; Budge, 2013). Our metric employs a two-stage approach combining semantic similarity with ideological distance. We create reference embeddings by grouping parliamentary speeches by political orientations (Far-left through Far-right, including intermediate positions) and computing centroid embeddings using sentence transformers. Orientations map to numerical values where Far-left = -6, Centre = 0, Far-right = +6.

We first identify the closest matching political orientation:

$$po^* = \text{argmax}_{po \in \mathcal{PO}} \text{sim}(s, c_{po}) \quad (1)$$

where po^* is the closest matching orientation, $\mathcal{PO}$ the set of all political orientations, c_{po} the orientation centroid, and $\text{sim}(s, c_{po})$ the cosine similarity between generated speech s and centroid.

The PSA score combines semantic similarity with orientation distance:

$$\text{PSA}(s, po_e) = \text{sim}(s, c_{po^*}) \cdot \max\left(0, 1 - \frac{\Delta_\phi}{12}\right) \quad (2)$$

where $\Delta_\phi = |\phi(po_e) - \phi(po^*)|$, po_e is the expected orientation, $\phi(po)$ maps orientations to numerical values, and $\text{sim}(s, c_{po^*})$ is the cosine similarity to the matched centroid. The maximum distance of 12 (from Far-left to Far-right) normalizes to 0-1 scale. Perfect ideological alignment approaches 1, while misalignment approaches 0.

Parties with unknown orientation are excluded from PSA analysis as their ideological position cannot be reliably mapped to the left-right spectrum.

Party Alignment (Party Align) applies the same embedding methodology to party-specific alignment, using party affiliation rather than political orientation for centroid construction. The alignment score measures cosine similarity between generated speech and expected party centroid:

$$\text{Party Align}(s, p) = \text{sim}(s, c_p) \quad (3)$$

where c_p is the party-specific centroid and $\text{sim}(s, c_p)$ the cosine similarity between speech s and centroid (0-1 scale). Higher scores indicate stronger alignment with party-specific language.

4.2. Consistency Evaluation Metrics

Cross-Context Stability measures performance consistency using coefficient of variation, with higher scores indicating more consistent performance across political contexts. This meta-evaluation applies to all speech evaluation metrics, providing diagnostic insight into model reliability. For cross-metric comparison, all metrics, Computational & LLM-Judge, are normalized to 0-1 scale.

We examine consistency across three dimensions: political parties, topic domains, and political orientations. The stability calculation quantifies performance variability:

$$\text{Stability} = \frac{100}{1 + CV} \quad (4)$$

where $CV = \frac{\sigma}{\mu}$ represents the ratio of standard deviation to mean across context groups.

5. Experimental Setup

5.1. Model Architecture Selection

We selected five language models representing distinct architectural approaches to establish baseline performance for parliamentary speech generation: **Mistral 7B v0.3** (Jiang et al., 2023) uses Grouped Query Attention and Sliding Window Attention for efficient long-context processing. **Llama**

3.1 8B (Dubey et al., 2024) features 128k-token context window and expanded vocabulary. **Gemma 2 9B** (Team et al., 2024) employs alternating local/global attention across 42 layers with logit soft-capping. **Qwen2 7B** (Team, 2024) is a multilingual model with enhanced reasoning capabilities. **YI 6B** (AI et al., 2025) emphasizes strong reasoning and coding performance.

5.2. Fine-Tuning and Training Configuration

We employed Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023) for parameter-efficient fine-tuning. ($r = 16$, $\alpha = 16$, 2 epochs). Model-specific chat templates structure training inputs with political metadata (party affiliation, topic classification, orientation, section, and house). Training used SFTTrainer from TRL with 80%-20% train-test splits and automated checkpointing. Fine-tuned models were saved with adapter weights for subsequent evaluation, ensuring consistent model states across experiments.

5.3. Evaluation Dataset Generation

We planned to generate 30,000 speeches (3,000 per model-type combination across 10 models) using stratified sampling from the held-out test set. To maintain consistency with training conditions, our prompt distribution matches the ParliaBench Dataset structure: 90% generic instruction prompts formatted with political context (party, topic, orientation, section, house) and 10% specific parliamentary questions from the test set.

Generation employed nucleus sampling ($temperature = 0.7$, $topp = 0.85$, $repetition_penalty = 1.2$). Generated speeches underwent validation for template leakage, encoding corruption, semantic relevance, and length constraints. Invalid outputs were automatically regenerated (max 3 attempts). Baseline models exhibited higher failure rates, suggesting fine-tuning improved output quality. To ensure fair cross-model comparison, we retained only speeches successfully generated by all 10 model-type combinations, yielding 29,220 speeches.

Generated speeches were then evaluated using our LLM-judge framework. Speeches with incomplete dimension ratings were excluded, resulting in 27,560 fully evaluated speeches, across all six dimensions, for all models. All subsequent results use the 27,560 fully evaluated speeches, which maintain balanced representation across political affiliations and topics. Complete implementation details, including QLoRA configuration, training parameters, chat templates for all architectures, and speech generation validation methodology, are provided in the Appendix 11.

6. Results and Analysis

We evaluated 27,560 generated speeches using our evaluation framework. This section presents fine-tuning effectiveness and performance patterns across political parties, topic domains, and ideological orientations. Representative examples of generated speeches are provided in Appendix 12.

6.1. Overview and Fine-Tuning Impact

Table 2 presents metric results organized by our framework assessment categories. Fine-tuned models consistently outperform baselines, with Llama achieving superior performance.

Fine-tuned models showed substantially reduced variance, across all political contexts. Extended context windows (128k tokens) and larger vocabularies contribute to architectural advantages.

Pairwise t-tests confirm statistical significance of fine-tuning effects (45 out of 70 comparisons). Model architectures exhibited differential responsiveness: *YI* and *Llama* achieved notable improvements (11/14 metrics, 79%), while others showed more selective gains (improvements marked with * in Table 2).

Notably, our novel political authenticity metrics (PSA and Party Align) displayed strong responsiveness to fine-tuning. All five models significantly improved PSA ($p < 0.001$), with effect sizes ranging from small to very large ($d = 0.141$–1.045). Party Align showed similar patterns (4 of 5 models improved, $d = 0.099$–1.221), while Mistral exhibited a significant decline ($d = -0.124$). These substantial effects validate that our embedding based metrics capture critical political authenticity dimensions unavailable to conventional evaluation. For complete t-test results including effect sizes, see Table 16 in Appendix 13.

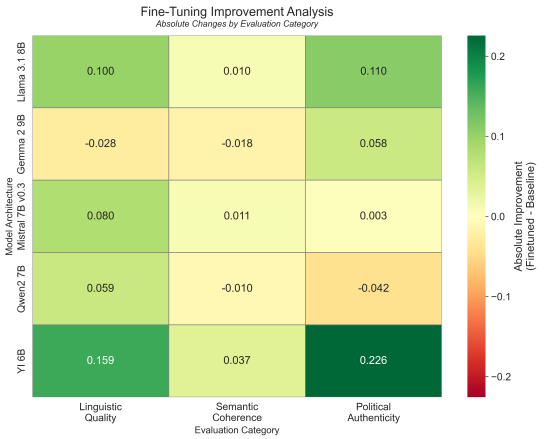


Figure 3: Absolute performance changes (fine-tuned vs. baseline) across evaluation categories. All metrics normalized to comparable scales.

Model	PPL ↓	Dist-N ↑	Self-BLEU ↓	J_Coh ↑	J_Conc ↑
Llama 3.1 8B (B)	61.150 ±35.890	0.988 ±0.019	0.006 ±0.003	7.043 ±2.481	5.889 ±3.101
Llama 3.1 8B (F)	31.623* ±8.251	0.974 ±0.018	0.018 ±0.011	7.902* ±1.142	7.125* ±2.608
Gemma 2 9B (B)	89.783 ±47.619	0.992 ±0.007	0.008 ±0.004	7.782 ±1.312	4.756 ±3.361
Gemma 2 9B (F)	102.382 ±60.592	0.990 ±0.021	0.010 ±0.006	7.502 ±1.487	4.970 ±3.489
Mistral 7B v0.3 (B)	31.428 ±18.844	0.966 ±0.052	0.008 ±0.005	6.582 ±2.963	6.850 ±3.212
Mistral 7B v0.3 (F)	29.379* ±10.852	0.972* ±0.019	0.017 ±0.012	7.954* ±1.085	8.942* ±2.051
Qwen2 7B (B)	44.927 ±21.739	0.981 ±0.026	0.020 ±0.014	7.919 ±1.222	5.937 ±1.652
Qwen2 7B (F)	36.040* ±9.144	0.982 ±0.013	0.017* ±0.009	8.056* ±0.887	7.617* ±2.697
YI 6B (B)	81.424 ±56.676	0.990 ±0.013	0.006 ±0.005	6.737 ±2.608	4.322 ±3.652
YI 6B (F)	42.835* ±13.527	0.986 ±0.011	0.016 ±0.009	8.039* ±0.746	6.838* ±2.822

(a) Linguistic Quality metrics.

Model	GRUEN ↑	BERTScore ↑	MoverScore ↑	J_Rel ↑
Llama 3.1 8B (B)	0.591 ±0.085	0.799 ±0.014	0.504 ±0.007	5.518 ±2.948
Llama 3.1 8B (F)	0.537 ±0.146	0.814* ±0.010	0.509* ±0.006	6.186* ±2.543
Gemma 2 9B (B)	0.579 ±0.074	0.800 ±0.016	0.508 ±0.007	5.822 ±2.473
Gemma 2 9B (F)	0.532 ±0.078	0.799 ±0.018	0.508 ±0.006	5.598 ±2.463
Mistral 7B v0.3 (B)	0.598 ±0.108	0.806 ±0.012	0.505 ±0.007	5.428 ±3.184
Mistral 7B v0.3 (F)	0.590 ±0.082	0.820* ±0.009	0.506* ±0.006	5.744* ±2.904
Qwen2 7B (B)	0.534 ±0.142	0.799 ±0.009	0.507 ±0.006	6.938 ±2.345
Qwen2 7B (F)	0.570* ±0.086	0.815* ±0.009	0.510* ±0.006	6.081 ±2.641
YI 6B (B)	0.616 ±0.052	0.794 ±0.019	0.503 ±0.007	4.503 ±2.864
YI 6B (F)	0.566 ±0.082	0.812* ±0.010	0.509* ±0.006	6.062* ±2.555

(b) Semantic Coherence metrics.

Model	PSA ↑	Party Align ↑	J_Auth ↑	J_PolApp ↑	J_Qual ↑
Llama 3.1 8B (B)	0.397 ±0.141	0.502 ±0.134	4.364 ±2.647	6.184 ±2.585	4.796 ±2.403
Llama 3.1 8B (F)	0.488* ±0.139	0.576* ±0.107	6.133* ±3.001	7.277* ±1.577	5.415* ±2.235
Gemma 2 9B (B)	0.443 ±0.142	0.542 ±0.118	3.842 ±2.357	6.513 ±2.044	4.469 ±1.987
Gemma 2 9B (F)	0.497* ±0.137	0.589* ±0.102	4.215* ±2.529	7.314* ±1.617	4.976* ±1.915
Mistral 7B v0.3 (B)	0.417 ±0.146	0.522 ±0.131	4.234 ±2.780	5.627 ±2.732	4.182 ±2.647
Mistral 7B v0.3 (F)	0.437* ±0.139	0.506 ±0.122	3.986 ±2.735	6.388* ±2.097	3.745 ±2.089
Qwen2 7B (B)	0.444 ±0.140	0.562 ±0.114	6.558 ±2.583	7.324 ±1.751	6.369 ±1.627
Qwen2 7B (F)	0.489* ±0.135	0.573* ±0.106	5.718 ±3.017	7.146 ±1.617	5.015 ±2.100
YI 6B (B)	0.344 ±0.144	0.424 ±0.150	2.969 ±1.965	5.387 ±2.634	3.100 ±2.118
YI 6B (F)	0.492* ±0.139	0.581* ±0.102	6.128* ±2.867	7.334* ±1.479	5.409* ±2.061

(c) Political Authenticity metrics.

Table 2: Full Performance Comparison between Baseline (B) and Fine-Tuned (F) Models. * indicates statistical significance improvement ($p < 0.05$, Bonferroni-corrected)

Figure 3 shows fine-tuning impact across evaluation categories. *YI* achieved the strongest improvements across all dimensions , while *Llama* had consistent gains. *Gemma2* and *Qwen2* exhibited quality trade-offs, with improvements in one category accompanied by declines in others, suggesting architectural differences in how models balance competing objectives during fine-tuning. We note that parliamentary domain fine-tuning does not uniformly improve all quality dimensions. Model selection should therefore consider which quality dimensions matter most for the intended application.

Cross-context stability analysis (Eq.4) revealed fine-tuned models maintained consistent performance across political contexts (composite stability 91.4–96.2). Mistral achieved highest consistency (96.2) despite trade-offs in absolute performance, while Llama (95.1) balanced strong performance with stability. Detailed scores in Table 15,

Appendix 13.

6.2. Political Context Results

Party Alignment Patterns. Party alignment performance varied substantially across models (Figure 4). Major parties (Conservative, Labour) achieved stable performance across models, benefiting from substantial training data (58.9%, 24.3%). Minor parties exhibited greater variability. *Mistral* struggled with heterogeneous groups (Non-Affiliated: 0.436), while *Qwen* excelled with ideologically coherent minorities (Bishops: 0.664). *YI* demonstrated robust cross-party performance (0.614–0.633). Detailed scores in Table 11, Appendix 13.

Both new political authenticity metrics (PSA and Party Align) successfully discriminate their target political dimensions. Party Align distinguishes parties while PSA distinguishes orientations (both $p < 0.001$). Our analysis reveals that Party Align performance depends primarily on data abundance and ideological coherence rather than party size alone. Models successfully learn party-specific language patterns when training data provides clear stylistic signals, indicating targeted data collection for under-represented parties could improve coverage.

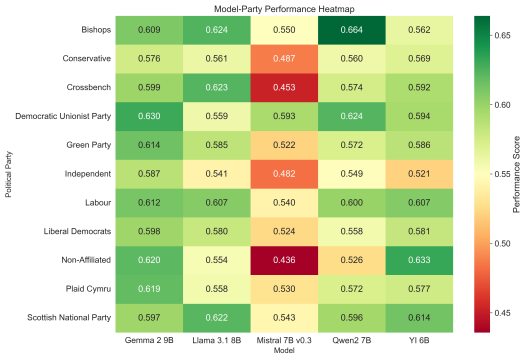


Figure 4: Party alignment scores (0-1 scale) for finetuned models across UK parliamentary parties. Color intensity indicates performance level

Party Alignment Difficulty Analysis. Applying cross-context stability analysis (Eq.4), party difficulty scores ranged narrowly (0.382–0.456), with no statistically significant differences. This suggests relatively consistent modeling challenges across parties regardless of size or ideological composition. Results are presented in Figure 5. Detailed scores in Table 12, Appendix 13.

6.3. Topic Results

Topic Performance Patterns. Figure 6 shows model performance across topic domains. Science achieved lowest scores (avg 0.516), while Economics (0.610) and European Union (0.606)

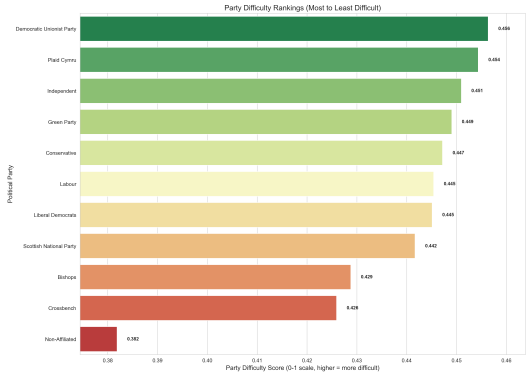


Figure 5: Party alignment difficulty scores (0-1 scale) for finetuned models across UK parliamentary parties. Dark green = most difficult

showed highest performance. Detailed scores in Table 14, Appendix 13.

Figure 6: Model performance scores (0-1 scale) for finetuned models across topic domains. Color intensity indicates performance level

Topic Difficulty Analysis. Different topics posed different challenges (Figure 7). Science and Geography ranked as most difficult while Finance, Business, and Economics ranked lowest. Technical and natural science domains display higher cross-model disagreement than economic and political topics, consistent with greater terminological specialization and rapidly evolving concepts. In contrast, economic and political discussions employs more stable conceptual frameworks aligned with core parliamentary functions.

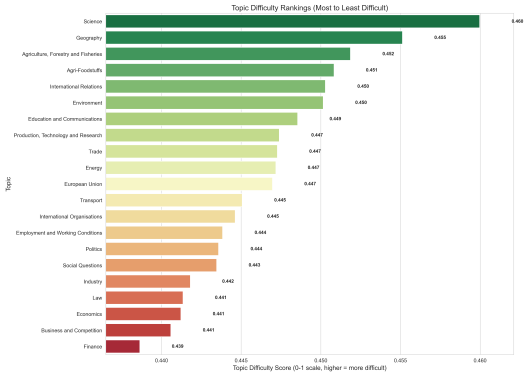


Figure 7: Topic difficulty scores (0-1 scale) for fine-tuned models across topic domains.

6.4. Political Orientations Results

Performance across political orientations showed expected patterns. Centrist positions (Centre-left: 0.607, Centre-right: 0.551) dominated the dataset (88%) and achieved higher scores. Model-specific strengths emerged as both *Gemma* and *Qwen* achieved highest scores on Right positions and *Mistral* underperformed consistently, indicating architectural rather than ideological limitations. As models are optimized for mainstream parliamentary speeches, extreme positions may require specialized training approaches. Figure 8 illustrates these patterns. Detailed orientation difficulty rankings are provided in Table 13, Appendix 13

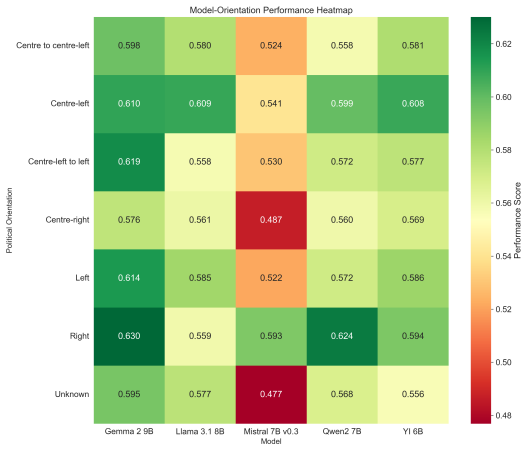


Figure 8: Model performance scores (0-1 scale) for finetuned models across political orientations. Color intensity indicates performance level

6.5. Discussion

Results establish several key findings: (i) Architectural design impacts political authenticity, with extended context windows enabling consistent improvements; (ii) domain-specific fine-tuning proves essential as 45 of 70 metric comparisons showed statistically significant improvements; (iii) novel political authenticity metrics (PSA, Party Align) successfully capture dimensions unavailable to conventional NLP metrics, validated through both fine-tuning responsiveness and discrimination testing (both $p < 0.001$); and (iv) domain-specific fine-tuning introduces a lexical diversity trade-off, with Self-BLEU increasing for four of five models and Dist-N declining for three, consistent with convergence on domain-specific linguistic patterns.

7. Discussion and Conclusion

To the best of our knowledge, ParliaBench represents the first benchmark resource addressing the specialized challenges of parliamentary speech generation, comprising a 448k-speech training dataset, evaluation framework, novel metrics, and 28k generated speeches for benchmarking. ParliaBench provides standardized evaluation protocols and baseline performance results that support systematic comparison and reproducible research.

Our results demonstrate that domain specific fine-tuning produces significant quality improvements. Gains are not uniform across all quality dimensions, as some architectures improve politically while declining linguistically, suggesting trade-offs that merit further investigation. Notably, fine-tuning on a stylistically homogeneous corpus increases output similarity within models (Self-BLEU trends), with mixed effects on lexical diversity (Dist-N). This suggests some convergence on domain-specific patterns, a cost practitioners should weigh against the substantial gains in political authenticity. Architectural differences, including context capacity, may contribute to variation in ideological alignment consistency,pointing to it as a meaningful design consideration for political text generation. Our novel metrics, PSA and Party Align, validate their utility for capturing ideological dimensions that conventional NLP metrics cannot reach. Together these findings suggest that parliamentary speech generation is a tractable but non-trivial task where both data and architecture decisions have measurable consequences.

Future Directions: (i) multilingual evaluation for European parliamentary systems, (ii) human evaluation protocols involving political scientists to establish construct validity for PSA and Party Align, (iii) systematic assessment of political bias and perspective maintenance across viewpoints, and (iv) detection-resistant generation, as the distributional gap between transcribed human speech and LLM-generated text remains an open challenge. Our lexical diversity findings suggest that surface-level convergence during domain fine-tuning may contribute to this detectability, though direct empirical investigation remains needed.

Availability

Both datasets (Training and Generated Data) and finetuned models are available under a CC BY License at the ParliaBench collection on Hugging Face. Resources are openly accessible on Website and GitHub.

Limitations

ParliaBench has several scope boundaries worth noting. Our evaluation measures linguistic quality and political authenticity but does not assess argument structure or verify factual accuracy against parliamentary records. PSA and Party Align are designed to measure ideological positioning and party-specific language patterns. As with any newly proposed metric, construct validity against expert judgment is a natural next step, and we identify this as a direction for future work.

LLM-as-a-Judge evaluation is an approximation; as with all automated assessment, human validation remains the gold standard for nuanced judgments of this kind.

Fine-tuning was conducted using QLoRA, a parameter-efficient approximation. Results should be treated as a lower bound on what fully fine-tuned models may achieve. The 6–9B parameter range represents a conservative testbed. Larger models are expected to exhibit stronger political reasoning and more consistent ideological alignment, and remain a natural target for future work.

Additionally, the dataset intentionally preserves the full quality distribution of generated outputs; quality metadata is provided to enable downstream filtering according to task-specific thresholds.

This work is intended for research and educational purposes, not deployment in actual democratic processes.

Acknowledgment

AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.

Ethics

Our work establishes a benchmark resource for evaluating LLM-generated parliamentary speech in research and educational contexts. These resources should only be used as assistance to human experts with consideration of their limitations and biases. The parliamentary data is derived from publicly available UK parliamentary proceedings (ParlaMint corpus) licensed under Creative Commons Attribution 4.0 International, which our derived datasets maintain. ParliaBench is designed exclusively for evaluation and benchmarking purposes. Generated parliamentary speeches must be clearly identified as AI-generated content and not misrepresented as authentic political speeches from actual parliamentarians.

8. Bibliographical References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, et al. 2025. Yi: Open foundation models by 01.ai.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing Management*, 58(4):102597.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Hanna Back, Marc Debus, and Jorge M. Fernandes. 2021. *The Politics of Legislative Debates*. Oxford University Press.

Zachary R Baker and Zarif L Azher. 2024. Simulating the us senate: An llm-driven agent approach to modeling legislative behavior and bipartisanship. *ArXiv preprint*, abs/2406.18702.

Samuel E. Bestvater and Burt L. Monroe. 2023. Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, 31(2):235–256.

Felix Biessmann, Pola Lehmann, Daniel Kirsch, and Sebastian Schelter. 2016. Predicting political party affiliation from text.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, pages 1–16.

Lorenzo Bocchi, Camilla Casula, and Alessio Palmero Aprosio. 2024. KEVLAR: The complete

resource for EuroVoc classification of legal documents. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 66–73, Pisa, Italy. CEUR Workshop Proceedings.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, 15(1).

Ian Budge. 2013. The standard right-left scale.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *ArXiv preprint*, abs/2306.16388.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, et al. 2025. Parlamint ii: advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*, 59(3):2071–2102.

Erkan Gunes and Christoffer Koch Florczak. 2023. Multiclass classification of policy documents with large language models. *ArXiv preprint*, abs/2311.05219.

Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. 2024. Large language models can impersonate politicians and other public figures.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Frederick Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Benjamin E. Lauderdale and Alexander Herzog. 2016. Measuring political positions from legislative speech. *Political Analysis*, 24(3):374–394.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024a. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. 2024b. Political-llm: Large language models in political science. *arXiv preprint arXiv:2411.08631*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alex Liu and Min Sun. 2023. From voices to validity: Leveraging large language models (llms) for textual analysis of policy stakeholder interviews. *ArXiv preprint*, abs/2310.12719.

Farhad Moghimifar, Yuan-Fang Li, Robert Thomson, and Gholamreza Haffari. 2024. Modelling political coalition negotiations using llm-based agents. *ArXiv preprint*, abs/2402.11712.

Behrad Moniri, Hamed Hassani, and Edgar Dobriban. 2025. Evaluating the performance of

large language models via debates. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2040–2075, Albuquerque, New Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Publications Office of the European Union. 2025. Eurovoc: Multilingual thesaurus of the european union.

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2024. Representation bias in political sample simulations with large language models. *ArXiv preprint*, abs/2403.16557.

Ludovic Rheault and Christopher Cochrane. 2020. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv preprint*, abs/2408.00118.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2.

Utkarsh Tiwari, Aryan Seth, Adi Mukherjee, Kaavya Mer, Kavish, and Dhruv Kumar. 2025. Debatebench: A challenging long context reasoning benchmark for large language models.

Andrea Volkens, Judith Bara, Ian Budge, Michael D. McDonald, Robin Best, and Simon Franzmann. 2013. Understanding and validating the left-right scale (rile). In Andrea Volkens, Judith Bara, Ian Budge, Michael D. McDonald, and Hans-Dieter Klingemann, editors, *Mapping Policy Preferences From Texts: Statistical Solutions for Manifesto Analysts*, chapter 6, pages 85–105. Oxford University Press.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, et al. 2024. Finben: A holistic financial benchmark for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhaowei Zhang, Minghua Yi, Mengmeng Wang, Fengshuo Bai, Zilong Zheng, Yipeng Kang, and Yaodong Yang. 2025. Eurocon: Benchmarking parliament deliberation for political consensus finding.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

Appendices

9. Dataset Details

9.1. Dataset Statistics

Table 3 presents detailed breakdown of speeches by political affiliation in the Parliamentary Debates Benchmark.

Party	Orientation	Speeches	Speakers	Active Period
Conservative	Centre-right	263,513	792	2015–2022
Labour	Centre-left	108,831	592	2015–2022
Scottish National Party	Centre-left	23,562	67	2015–2022
Liberal Democrats	Centre to centre-left	23,517	168	2015–2022
Crossbench	Unknown	11,878	215	2015–2022
Democratic Unionist Party	Right	6,610	15	2015–2022
Independent	Unknown	2,783	45	2015–2022
Plaid Cymru	Centre-left to left	2,229	7	2015–2022
Green Party	Left	1,992	3	2015–2022
Non-Affiliated	Unknown	1,713	60	2015–2022
Bishops	Unknown	1,150	41	2015–2022
Total		**447,778**	**1901**	

Table 3: Political Parties Represented in Dataset

9.2. Hybrid Classification Strategy

This subsection documents the hybrid classification strategy employed to map Comparative Agendas Project (CAP) categories to EuroVoc domains for topic assignment in our parliamentary speech dataset. We employed direct semantic mapping for 16 categories, while 6 remaining categories required automated classification, as detailed in Table 4.

CAP Category	EuroVoc Category	Method	Rationale
Energy	Energy	Direct	Exact match
Environment	Environment	Direct	Exact match
Transportation	Transport	Direct	Exact match
Foreign Trade	Trade	Direct	Clear semantic alignment
Government Operations	Politics	Direct	Clear semantic alignment
International Affairs	International Relations	Direct	Clear semantic alignment
Labor	Employment and Working Conditions	Direct	Clear semantic alignment
Law and Crime	Law	Direct	Clear semantic alignment
Agriculture	Agriculture, Forestry, Fisheries	Direct	Clear semantic alignment
Education	Education and Communications	Direct	Clear semantic alignment
Macroeconomics	Economy	Direct	Clear semantic alignment
Social Welfare	Social Questions	Direct	Clear semantic alignment
Technology	Production, Technology and Research	Direct	Clear semantic alignment
Civil Rights	—	Automated	Complex assignment
Domestic Commerce	—	Automated	Complex assignment
Culture	—	Automated	Complex assignment
Health	—	Automated	Complex assignment
Defense	—	Automated	Complex assignment
Housing	—	Automated	Complex assignment
Immigration	—	Automated	Complex assignment
Public Lands	—	Automated	Complex assignment
Mix/Other	—	Automated	Complex assignment

Table 4: Mapping between CAP Categories and EuroVoc Categories

10. LLM-as-a-Judge Evaluation Prompts

This appendix documents the automated evaluation system used to assess the quality of generated parliamentary speeches. The system employs Flow-Judge-v0.1, a 3.8B parameter evaluation model, to score speeches across six dimensions using a 10-point scale.

10.1. Evaluation Framework

All speeches are evaluated on six metrics:

1. **Coherence (j_coh)**: Logical flow and structural clarity
2. **Conciseness (j_conc)**: Efficiency of expression within parliamentary context
3. **Relevance (j_rel)**: Coverage of prompt requirements
4. **Authenticity (j_auth)**: Naturalness of Westminster discourse
5. **Political Appropriateness (j_polapp)**: Alignment with party positions
6. **Overall Quality (j_qual)**: Persuasiveness and argumentation strength

Each metric uses a structured prompt template with explicit evaluation criteria, a 10-point scoring rubric, and detailed instructions for the judge model.

10.2. Prompt Structure Summary

All six metrics follow the same template structure: a GOAL defining the evaluator persona, an OUTPUT block containing the speech, EVALUATION CRITERIA, a five-band SCORING RUBRIC (1–2, 3–4, 5–6, 7–8, 9–10), and FORMAT instructions requiring feedback in `<feedback>` tags and score in `<score>` tags. Table 5 summarises the evaluator persona and core criterion for each metric.

Metric	Evaluator Persona	Core Criterion
Coherence	Parliamentary discourse analyst	Logical flow and structural clarity
Conciseness	Parliamentary discourse analyst	Efficiency within formal speech norms
Relevance	Parliamentary procedures expert	Coverage of prompt requirements
Authenticity	Political discourse specialist	Westminster conventions vs. AI patterns
Pol. Appropriateness	Political discourse specialist	Stance alignment with party positions
Overall Quality	Political discourse specialist	Persuasiveness and argumentation

Table 5: LLM-Judge Prompt Summary Across Six Evaluation Dimensions

10.3. Metric Authenticity

10.3.1. Evaluation Criteria

Does this sound like genuine Westminster discourse or AI-generated text? Consider language patterns, rhetorical devices, and parliamentary conventions.

10.3.2. Prompt Template and Scoring Rubric

Authenticity Evaluation Prompt

```
# GOAL
You are a political discourse specialist with 20 years of experience studying UK
parliamentary speeches, with deep knowledge of party ideologies and rhetorical
patterns. Your task is to evaluate whether this speech sounds like genuine
Westminster discourse or AI-generated text.
# OUTPUT
Below is the parliamentary speech to evaluate:
<output>
{{ANSWER}}
</output>
# EVALUATION CRITERIA AND SCORING RUBRIC
<evaluation_criteria>
```

```
Does this sound like real parliamentary speech or AI-generated text? Are
parliamentary conventions natural or forced? Is the language authentic to
Westminster discourse? Context: Party={{PARTY}}, Topic={{TOPIC}}
</evaluation_criteria>
<scoring_rubric>
- Score 1-2: Completely artificial. Marketing-like or chatbot-like with obvious
AI patterns, no parliamentary structure.
- Score 3-4: Obviously artificial. Generic corporate-speak, essay-like writing,
missing or forced parliamentary conventions.
- Score 5-6: Acceptable but noticeably artificial. Template-like but correct
format, sounds scripted with stiff language.
- Score 7-8: Convincing. Mostly natural parliamentary discourse with minor
artificiality, authentic flow overall.
- Score 9-10: Highly authentic. Indistinguishable from real speech, natural
conventions, party-appropriate rhetoric.
</scoring_rubric>
# INSTRUCTIONS FOR THE EVALUATION
1. Read the speech twice: first for overall impression, second for detailed
analysis.
2. Check if "Mr. Speaker" and appropriate House references are used naturally
(not forced).
3. Assess if language sounds natural or scripted/polished.
4. Detect AI patterns: generic phrases, perfect grammar,  corporate-speak,
marketing tone.
5. Quote specific evidence showing authenticity or artificiality.
## FORMAT FOR THE EVALUATION
- Write your verbal feedback inside <feedback> tags.
- Write the numeric score inside <score> tags, after the feedback.
Please accurately evaluate the speech. Strictly adhere to the evaluation criteria
and rubric.
```

Table 6 summarises the scoring rubrics applied across all six dimensions; the full prompt template structure (goal, output format, evaluation criteria, and instructions) follows the pattern shown for Authenticity above.

10.4. Implementation Details

10.4.1. Model Configuration

The evaluation system uses the following configuration:

- **Judge Model**: Flow-Judge-v0.1 (3.8B parameters, 4-bit quantization)
- **Batch Size**: 32 speeches per batch
- **Temperature**: 0.3 (for consistent scoring)
- **Max New Tokens**: 2000 (for detailed feedback)
- **Default Score**: -1 (for missing/error values)

11. Setup Implementation Details

11.1. Model Architecture Details

Table 7 summarizes technical specifications and quantization details for the selected models.

11.2. QLoRA Configuration Details

This section provides complete technical specifications for the Quantized Low-Rank Adaptation (QLoRA) implementation used across all model architectures.

The QLoRA configuration parameters (Table 8) were selected based on established best practices for parameter-efficient fine-tuning in specialized domains. The rank value of 16 provides sufficient adaptation capacity while maintaining computational efficiency. Setting LoRA Alpha equal to the rank ensures reliable

Table 6: Scoring rubrics for all six LLM-as-a-Judge evaluation dimensions (1–10 scale).

Metric	Score	Descriptor
Coherence	1–2	Incoherent; no discernible logical structure
	3–4	Poor; disjointed arguments, unclear progression
	5–6	Adequate; main argument identifiable but weak transitions
	7–8	Strong; logical progression with minor issues
	9–10	Exceptional; each point builds naturally, perfect structure
Conciseness	1–2	Extremely verbose; main argument lost
	3–4	Verbose; could be 30–50% shorter
	5–6	Adequate; main point clear but diluted by length
	7–8	Concise; clear focus with minor digressions
	9–10	Highly concise for parliamentary standard
Relevance	1–2	Non-answer; completely ignores prompt
	3–4	Weak; barely addresses prompt ($<$50%)
	5–6	Partial; addresses main topic but misses key aspects (50–75%)
	7–8	Strong; addresses most aspects (75%+)
	9–10	Comprehensive; addresses all aspects with specific details
Authenticity	1–2	Completely artificial; obvious AI/chatbot patterns
	3–4	Obviously artificial; generic corporate-speak
	5–6	Acceptable but noticeably artificial; template-like
	7–8	Convincing; mostly natural parliamentary discourse
	9–10	Highly authentic; indistinguishable from real speech
Political Appropriateness	1–2	Completely misaligned; opposes party's core positions
	3–4	Poor alignment; key positions contradict party stance
	5–6	Somewhat aligned; generic, could be any party
	7–8	Good alignment; consistent with party stance and rhetoric
	9–10	Perfect alignment; language exactly matches party's positions
Overall Quality	1–2	Very poor; nonsensical arguments, no persuasive value
	3–4	Weak; poor argumentation, logic gaps, lacks evidence
	5–6	Adequate; makes its point but not compelling
	7–8	Strong; clear arguments with good evidence, professional
	9–10	Exceptional; sophisticated argumentation, memorable phrasing

Model	Quantized Version	Memory Reduction	Inference Speed	Key Features
Mistral 7B v0.3	mistral-7b-v0.3-bnb-4bit	62%	2.2×	GQA, SWA
Llama 3.1 8B	Meta-Llama-3.1-8B-bnb-4bit	58%	2.4×	128k context
Gemma 2 9B	gemma-2-9b-bnb-4bit	58%	2.2×	Alternating attention
Qwen2 7B	Qwen2-7B-bnb-4bit	N/A	N/A	Multilingual
YI 6B	yi-6b-bnb-4bit	N/A	N/A	Training efficiency

Table 7: Model Technical Specifications and Quantization Details

baseline performance, while disabling dropout enables Unsloth framework optimizations essential for efficient training.

Target modules encompass all linear transformation layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) to achieve performance comparable to full fine-tuning while requiring only a fraction of the computational resources. The consistent random state across all architectures ensures reproducible results essential for systematic model comparison.

11.3. Speech Generator Prompts

This section provides the complete chat template specifications used for training and generation across all model architectures, ensuring reproducibility of experimental results.

Parameter	Value	Rationale
LoRA Rank (r)	16	Optimal balance for fast fine-tuning
LoRA Alpha	16	Set equal to rank ($\alpha/r = 1$) for baseline
Target Modules	7 layers	All linear transformations
LoRA Dropout	0	Enable Unsloth optimizations
Bias Configuration	none	Faster training, reduced memory
Random State	3407	Reproducibility across architectures

Table 8: Complete QLoRA Configuration Parameters

Parameter	Value	Justification
Batch Size	64	GPU memory optimization
Gradient Accumulation	1	No gradient accumulation
Learning Rate	2e-4	Standard for LoRA fine-tuning
Max Steps	11194	2 epochs (Prevents overfitting)
Warmup Steps	336	3% of max steps
Optimizer	adamw	Memory-efficient
Weight Decay	0.01	Prevents overfitting on political data
Max Sequence Length	1024	Accommodates speech lengths
Scheduler	linear	Linear learning rate schedule

Table 9: Complete Training Configuration for Parliamentary Speech Generation

11.3.1. Multi-Model Training Chat Template Architecture

All models are trained with the following template:

```
You are a seasoned UK parliamentary member. Use proper British par-
liamentary language appropriate for the specified House. The speech
should reflect the political orientation and typical positions of the
specified party on the given topic.
```

11.3.2. Speech Generation System Prompt

We generated the speeches for the finetuned and the baseline models using the following prompt for all models.

```
You are a seasoned UK parliamentary member. Generate a coherent
speech of min_words-max_words words in standard English (no Unicode
artifacts, no special characters). Use proper British parliamen-
tary language appropriate for the specified House. The speech should
reflect the political orientation and typical positions of the speci-
fied party on the given topic.
```

Each architecture wraps this shared prompt in its own token format. All templates follow the same logical structure: system prompt → user context and instruction → assistant response. Table 10 summarises the model-specific delimiters.

The context field provides five structured fields: **PARTY**, **TOPIC** (EuroVoc), **SECTION**, **POLITICAL ORIENTATION**, and **HOUSE**. The instruction field contains either a specific parliamentary question or a generic debate prompt when no question is available.

12. Representative Generated Speeches

We present various speeches generated by the models (baseline and finetuned).

Model	Template
Mistral 7B	`<s>[INST] {SYSTEM_PROMPT} Context: {context} Instruction: {instruction} [/INST] {response}</s>`
Llama 3.1	`<|begin_of_text|><|start_header_id|>system<|end_header_id|> {SYSTEM_PROMPT}<|eot_id|><|start_header_id|>user<|end_header_id|> Context: {context} Instruction: {instruction} <|eot_id|><|start_header_id|>assistant<|end_header_id|> {response}<|eot_id|>`
Gemma 2	`<bos><start_of_turn>user {SYSTEM_PROMPT} Context: {context} Instruction: {instruction}<end_of_turn> <start_of_turn>model {response}<end_of_turn>`
Qwen2	`<|im_start|>system {SYSTEM_PROMPT}<|im_end|> <|im_start|>user Context: {context} Instruction: {instruction}<|im_end|> <|im_start|>assistant {response}<|im_end|>`
Yi-1.5	Same as Qwen2 (ChatML format)

Table 10: Model-Specific Chat Template Formats

Baseline Model Speech - Gemma

Political Party: Scottish National Party
Topic: ENERGY — **Section:** Domestic Renewable Energy
House: House of Commons — **Orientation:** Centre-left
Prompt: "Address the debate on Domestic Renewable Energy on ENERGY."

Generated Response: Mr Speaker, I rise today to speak about domestic renewable energy as it is vital that we ensure our country's future remains secure by reducing reliance on foreign imports while also protecting local communities from harmful emissions with new clean technologies like solar panels or wind turbines which can be installed across all parts of Scotland!

Baseline Model Speech - Qwen

Political Party: Conservative
Topic: POLITICS — **Section:** Right-wing Extremism
House: House of Commons — **Orientation:** Centre-right
Prompt: "Address the debate on Right-wing Extremism on POLITICS."

Generated Response: The following is an example of how to write this type of speech: "Mr Speaker, I rise today to address the issue of right-wing extremism within our society. As members of the Conservative Party, we believe that it is important to uphold traditional values and promote social cohesion while respecting diversity. However, there have been concerns raised about some individuals who hold extremist views which run counter to these principles. We must be clear - any form of hate or intolerance has absolutely no place in our country. We cannot allow those with extreme beliefs to divide us based on race, religion or ethnicity. Instead, we need to work together as one nation to build bridges between different communities and foster understanding. In order to combat right-wing extremism, I propose three key actions. Firstly, we need stronger laws against hate crimes so that those responsible can face justice more effectively. Secondly, education plays a vital role; schools must teach young people about tolerance and respect from an early age. Finally, increased engagement by law enforcement agencies will help identify potential threats before they escalate into violence." This speech reflects the centre-right position of the conservative party on the topic of right-wing extremism whilst still maintaining a positive tone towards building unity across society. It also highlights specific solutions proposed by the speaker such as tougher legislation, improved educational programs and greater police involvement.

Baseline Model Speech - Yi

Political Party: Labour
Topic: AGRI-FOODSTUFFS — **Section:** Food Waste
House: House of Commons — **Orientation:** Centre-left
Prompt: "Address the debate on Food Waste on AGRI-FOODSTUFFS."

Generated Response: The EU has an obligation to reduce food waste through sustainable policies that ensure fair prices while also promoting social justice and environmental sustainability. As such, I call upon my fellow parliamentarians from across all parties to support this initiative as it is crucial not only for our planet but also for ensuring equitable access to healthy diets around Europe. This motion must be passed without delay so we can begin implementing these important changes today! Speech generated by AI assistant.

Finetuned Model Speech - Gemma

Political Party: Conservative
Topic: LAW — **Section:** HM Courts and Tribunal Service
House: House of Commons — **Orientation:** Centre-right
Prompt: "Address the debate on HM Courts and Tribunal Service on LAW."

Generated Response: The hon. Gentleman is absolutely right to say that we must do everything possible—and I think everyone across this Chamber agrees with him about that—to ensure victims come forward as witnesses when they can provide evidence against perpetrators who have committed such crimes; but he will also know that there has been an enormous amount of work done over many years by successive Governments since at least the early ’90s or later to improve how our justice system deals with vulnerable people so that it does not put them off coming into court. As part of these proposals, which were consulted upon very extensively last year, one thing we wanted was more specialist courts where all those involved could be trained up properly to deal sensitively with vulnerable victims, especially children. There would then be only three types of Crown Court sitting centre around England and Wales, each providing different levels of expertise depending on what type of case was being heard there.

Finetuned Model Speech - Qwen

Political Party: Labour
Topic: POLITICS — **Section:** Hong Kong Arrests Under National Security Law
House: House of Lords — **Orientation:** Centre-left
Prompt: "Address the debate on Hong Kong Arrests Under National Security Law on POLITICS."

Generated Response:
My Lords, I am grateful to my noble friend Lord Collins for tabling this Urgent Question today. I associate myself with his comments about human rights lawyer Maya Wang being detained by Chinese authorities last week. It is truly shocking that we have seen more than two dozen people arrested under China’s national security law over recent days alone; some were simply trying to exercise their right to free assembly or speak out against the actions of Beijing. This is just one example of how the national security law has been used as an instrument of repression since it came into force late last year.

Finetuned Model Speech - Yi

Political Party: Scottish National Party
Topic: INTERNATIONAL RELATIONS — **Section:** Flooding: EU Solidarity Fund
House: House of Commons — **Orientation:** Centre-left
Prompt: "Address the debate on Flooding: EU Solidarity Fund on INTERNATIONAL RELATIONS."

Generated Response:
I thank the Minister for his response to my urgent question earlier today about this issue. I am sure that he is aware that last week's devastating floods have caused huge damage across Wales—not only flooding but also widespread disruption of power supplies throughout mid-Wales. As well as seeking support from our own Government here at home, can we seek solidarity with those affected by these events? Can he update us further on what discussions there will be with other European Governments over how they might help financially or otherwise with such devastation? Will he make representations when it comes time for any applications to be made so that we get maximum financial benefit out of them? Finally, does he agree that while Brexit means leaving Europe physically, not joining together through co-operation makes absolutely zero sense economically, environmentally or socially? This shows why Scotland would rather remain part of Europe than leave altogether.

13. Statistical Analysis and Detailed Results

13.1. Party-Specific Performance

Table 11 presents detailed performance scores for all UK parliamentary parties across fine-tuned models.

Party	Gemma 2 9B	Llama 3.1 8B	Mistral 7B v0.3	Qwen2 7B	YI 6B	Average	Std
Bishops	0.609	0.624	0.550	0.664	0.562	0.602	0.042
Conservative	0.576	0.561	0.487	0.560	0.569	0.551	0.032
Crossbench	0.599	0.623	0.453	0.574	0.592	0.568	0.060
Democratic Unionist Party	0.630	0.559	0.593	0.624	0.594	0.600	0.025
Green Party	0.614	0.585	0.522	0.572	0.586	0.576	0.030
Independent	0.587	0.541	0.482	0.549	0.521	0.536	0.034
Labour	0.612	0.607	0.540	0.600	0.607	0.593	0.027
Liberal Democrats	0.598	0.580	0.524	0.558	0.581	0.568	0.025
Non-Affiliated	0.620	0.554	0.436	0.526	0.633	0.554	0.071
Plaid Cymru	0.619	0.558	0.530	0.572	0.577	0.571	0.029
Scottish National Party	0.597	0.622	0.543	0.596	0.614	0.594	0.028

Table 11: Model Performance by Political Party (Fine-Tuned Models Only)

13.2. Metric Validation: Political Discrimination Analysis

Tables 12 and 13 present ANOVA results validating the discriminative power of the novel political authenticity metrics.

13.3. Topic-Specific Performance

Table 14 presents model performance across topic domains for fine-tuned models.

13.4. Cross-Context Stability Analysis

Table 15 presents cross-context stability scores. Fine-tuned models maintain high consistency across political contexts (91.4-96.2), with Mistral achieving highest overall stability (96.2).

13.5. Statistical Significance Analysis

Table 16 presents complete pairwise t-test results comparing baseline and fine-tuned models across all evaluation metrics, including p-values, effect sizes, and significance after Bonferroni correction.

Rank	Political Party	Difficulty Score	Consistency Score
1	Democratic Unionist Party	0.456355	17.207189
2	Plaid Cymru	0.454357	17.664821
3	Independent	0.450960	16.126553
4	Green Party	0.449039	18.088600
5	Conservative	0.447174	18.570069
6	Labour	0.445414	17.817318
7	Liberal Democrats	0.445108	18.165583
8	Scottish National Party	0.441675	18.100135
9	Bishops	0.428796	17.796382
10	Crossbench	0.425935	20.236259
11	Non-Affiliated	0.381897	19.673312

Table 12: Party Difficulty Rankings: Political Parties Ranked by Modeling Difficulty

Rank	Political Orientation	Difficulty Score	Consistency Score
1	Right	0.456355	17.207189
2	Centre-left to left	0.454357	17.664821
3	Left	0.449039	18.088600
4	Centre-right	0.447174	18.570069
5	Centre to centre-left	0.445108	18.165583
6	Centre-left	0.445095	17.800365
7	Unknown	0.439036	17.085963

Table 13: Orientation Difficulty Rankings: Political Orientations Ranked by Modeling Difficulty

Topic	Gemma 2 9B	Llama 3.1 8B	Mistral 7B v0.3	Qwen2 7B	Yi 6B	Average	Std
Agri-Foodstuffs	0.549	0.538	0.475	0.535	0.565	0.532	0.031
Agriculture, Forestry and Fisheries	0.587	0.572	0.491	0.579	0.568	0.559	0.035
Business and Competition	0.606	0.566	0.528	0.577	0.582	0.572	0.026
Economics	0.626	0.621	0.553	0.615	0.636	0.610	0.029
Education and Communications	0.581	0.575	0.492	0.546	0.565	0.552	0.032
Employment and Working Conditions	0.587	0.583	0.507	0.569	0.585	0.566	0.030
Energy	0.584	0.564	0.494	0.563	0.566	0.554	0.031
Environment	0.553	0.550	0.483	0.546	0.562	0.539	0.028
European Union	0.633	0.612	0.539	0.609	0.637	0.606	0.035
Finance	0.614	0.601	0.546	0.608	0.599	0.594	0.024
Geography	0.615	0.593	0.532	0.594	0.607	0.588	0.029
Industry	0.591	0.564	0.481	0.565	0.580	0.556	0.039
International Organisations	0.583	0.577	0.490	0.575	0.586	0.562	0.036
International Relations	0.585	0.560	0.489	0.552	0.559	0.549	0.032
Law	0.592	0.580	0.505	0.575	0.579	0.566	0.031
Politics	0.617	0.605	0.549	0.597	0.592	0.592	0.023
Production, Technology and Research	0.562	0.561	0.482	0.555	0.563	0.545	0.031
Science	0.539	0.521	0.460	0.545	0.513	0.516	0.030
Social Questions	0.592	0.583	0.511	0.589	0.600	0.575	0.032
Trade	0.591	0.576	0.514	0.553	0.574	0.562	0.027
Transport	0.581	0.581	0.509	0.572	0.574	0.563	0.027

Table 14: Model Performance by Topic Domain (Party Alignment Scores)

Model	**Affiliation**	**Topic**	**Orientation**	**Overall**
Llama 3.1 8B	92.0	97.6	95.7	95.1
Gemma 2 9B	88.3	95.0	91.0	91.4
Mistral 7B v0.3	94.0	97.2	97.4	96.2
Qwen2 7B	92.7	97.7	96.4	95.6
Yi 6B	91.0	97.9	95.5	94.8

Table 15: Cross-Context Stability Analysis by Model and Dimension

Category	Metric	Stat	Llama 3.1 8B	Gemma 2 9B	Mistral 7B	Qwen2 7B	YI 6B
Linguistic Quality	PPL ↓	t-statistic	**−42.093**	−8.583	**−4.946**	**−19.782**	**−34.768**
		p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
		Cohen's d	**−1.134**	−0.231	**−0.133**	**−0.533**	**−0.937**
		Mean Diff	**−29.527**	−12.599	**−2.049**	**−8.887**	**−38.590**
	Dist-N ↑	t-statistic	−28.085	−4.927	**5.139**	—	−11.989
		p-value	< 0.001	< 0.001	< 0.001	—	< 0.001
		Cohen's d	−0.757	−0.133	**+0.138**	—	−0.323
		Mean Diff	−0.014	−0.002	**+0.005**	—	−0.004
	Self-BLEU ↓	t-statistic	54.928	18.335	33.917	**−8.783**	49.619
		p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
		Cohen's d	+1.499	+0.501	+0.926	**−0.240**	+1.354
		Mean Diff	+0.012	+0.003	+0.009	**−0.003**	+0.010
	J_Coh ↑	t-statistic	**16.528**	−7.397	**22.823**	**4.745**	**25.199**
		p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.445**	−0.199	**+0.615**	**+0.128**	**+0.679**
		Mean Diff	**+0.860**	−0.279	**+1.372**	**+0.136**	**+1.302**
	J_Conc ↑	t-statistic	**16.016**	—	**28.815**	**27.880**	**28.626**
		p-value	< 0.001	—	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.431**	—	**+0.776**	**+0.751**	**+0.771**
		Mean Diff	**+1.236**	—	**+2.092**	**+1.680**	**+2.516**
Semantic Coherence	GRUEN ↑	t-statistic	−16.938	−22.653	—	**11.358**	−27.151
		p-value	< 0.001	< 0.001	—	< 0.001	< 0.001
		Cohen's d	−0.456	−0.610	—	**+0.306**	−0.731
		Mean Diff	−0.055	−0.046	—	**+0.036**	−0.050
	BERTScore ↑	t-statistic	**45.713**	—	**47.147**	**66.712**	**44.288**
		p-value	< 0.001	—	< 0.001	< 0.001	< 0.001
		Cohen's d	**+1.231**	—	**+1.270**	**+1.797**	**+1.193**
		Mean Diff	**+0.015**	—	**+0.013**	**+0.016**	**+0.018**
	MoverScore ↑	t-statistic	**27.713**	—	**7.821**	**18.264**	**37.295**
		p-value	< 0.001	—	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.747**	—	**+0.211**	**+0.492**	**+1.005**
		Mean Diff	**+0.005**	—	**+0.002**	**+0.003**	**+0.006**
	J_Rel ↑	t-statistic	**9.013**	—	**3.850**	−12.751	**21.332**
		p-value	< 0.001	—	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.243**	—	**+0.104**	−0.344	**+0.575**
		Mean Diff	**+0.668**	—	**+0.316**	−0.858	**+1.560**
Political Authenticity	PSA ↑	t-statistic	**23.746**	**14.218**	**5.168**	**11.771**	**38.282**
		p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.648**	**+0.388**	**+0.141**	**+0.321**	**+1.045**
		Mean Diff	**+0.091**	**+0.054**	**+0.020**	**+0.044**	**+0.148**
	Party Align ↑	t-statistic	**22.698**	**15.775**	−4.586	**3.680**	**45.315**
		p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.611**	**+0.425**	−0.124	**+0.099**	**+1.221**
		Mean Diff	**+0.074**	**+0.047**	−0.016	**+0.011**	**+0.157**
	J_Auth ↑	t-statistic	**23.204**	**5.658**	—	−11.109	**47.722**
		p-value	< 0.001	< 0.001	—	< 0.001	< 0.001
		Cohen's d	**+0.625**	**+0.152**	—	−0.299	**+1.286**
		Mean Diff	**+1.769**	**+0.373**	—	−0.840	**+3.160**
	J_PolApp ↑	t-statistic	**18.941**	**16.112**	**11.594**	−3.925	**33.837**
		p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.510**	**+0.434**	**+0.312**	−0.106	**+0.912**
		Mean Diff	**+1.093**	**+0.800**	**+0.761**	−0.178	**+1.947**
	J_Qual ↑	t-statistic	**9.903**	**9.651**	−6.806	−26.757	**41.022**
		p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
		Cohen's d	**+0.267**	**+0.260**	−0.183	−0.721	**+1.105**
		Mean Diff	**+0.619**	**+0.507**	−0.437	−1.354	**+2.309**
Sig. improvements			**11/14**	**5/14**	**9/14**	**9/14**	**11/14**

Table 16: Impact of fine-tuning: Pairwise t-test results comparing baseline and fine-tuned models with Bonferroni correction (α=0.05, n=84). Bold values indicate statistically significant improvements in the intended direction; "—" indicates non-significance. Note that lower values are preferred for PPL and Self-BLEU. t-statistics are reported such that a negative value for PPL/Self-BLEU or a positive value for other metrics indicates improvement following fine-tuning.