



NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF COMPUTER SCIENCE

Evaluating the Synthetic Generation of Political Speech Using Large Language Models

DIPLOMA THESIS

of

ARGYRO TSIPI

Supervisor: Panayiotis Tsanakas
Professor NTUA

Athens, October 2025



NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF COMPUTER SCIENCE

Evaluating the Synthetic Generation of Political Speech Using Large Language Models

DIPLOMA THESIS

of

ARGYRO TSIPI

Supervisor: Panayiotis Tsanakas
Professor NTUA

Approved by the examination committee on 30th October 2025.

(Signature)

(Signature)

(Signature)

.....
Panayiotis Tsanakas
Professor NTUA

.....
Andreas Stafylopatis
Professor Emeritus NTUA

.....
Georgios Matsopoulos
Professor NTUA

Athens, October 2025



Copyright © – All rights reserved.

Argyro Tsipi, 2025.

The copying, storage and distribution of this diploma thesis, exall or part of it, is prohibited for commercial purposes. Reprinting, storage and distribution for non - profit, educational or of a research nature is allowed, provided that the source is indicated and that this message is retained.

The content of this thesis does not necessarily reflect the views of the Department, the Supervisor, or the committee that approved it.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....

Argyro Tsipi

Graduate of School of Electrical and Computer Engineering, National Technical University of Athens

Περίληψη

Η παραγωγή κοινοβουλευτικών λόγων παρουσιάζει συγκεκριμένες προκλήσεις για τα μεγάλα γλωσσικά μοντέλα πέρα από τις τυπικές εργασίες παραγωγής κειμένου. Σε αντίθεση με τη γενική παραγωγή κειμένου, οι κοινοβουλευτικές ομιλίες απαιτούν όχι μόνο γλωσσική ποιότητα αλλά και πολιτική αυθεντικότητα και ιδεολογική συνέπεια. Τα τρέχοντα γλωσσικά μοντέλα δεν διαθέτουν εξειδικευμένη εκπαίδευση για κοινοβουλευτικά πλαίσια, και οι υπάρχουσες μέθοδοι αξιολόγησης επικεντρώνονται σε τυπικές μετρήσεις NLP παρά στην πολιτική αυθεντικότητα. Για να αντιμετωπίσουμε αυτό, παρουσιάζουμε ένα πλαίσιο για την παραγωγή κοινοβουλευτικών λόγων. Κατασκευάσαμε και προεπεξεργαστήκαμε ένα σύνολο δεδομένων ομιλιών από το Κοινοβούλιο του Ηνωμένου Βασιλείου Parlamint GB για να επιτρέψουμε τη συστηματική εκπαίδευση μοντέλων. Εισάγουμε ένα ολοκληρωμένο πλαίσιο αξιολόγησης που συνδυάζει υπολογιστικές μετρήσεις με αξιολογήσεις LLM-as-a-judge για τη μέτρηση της ποιότητας παραγωγής σε τρεις διαστάσεις: γλωσσική ποιότητα, σημασιολογική συνοχή και πολιτική αυθεντικότητα. Για τη γλωσσική ποιότητα χρησιμοποιήσαμε μετρήσεις όπως Perplexity, Self-BLEU, BERTScore, GRUEN Score, MOVER Score και Distinct-n. Προτείνουμε δύο νέες μετρήσεις βασισμένες σε embeddings, το Political Spectrum Alignment και το Party Alignment, για να ποσοτικοποιήσουμε την ιδεολογική τοποθέτηση. Επιπλέον, χρησιμοποιήσαμε την προσέγγιση LLM-as-a-judge για να αξιολογήσουμε έξι διαστάσεις: συνοπτικότητα, συνοχή, αυθεντικότητα, πολιτική καταλληλότητα, συνολική ποιότητα και συνάφεια. Εκπαιδεύσαμε με fine-tuning πέντε μεγάλα γλωσσικά μοντέλα (Mistral, Gemma, Qwen, Llama, Yi) χρησιμοποιώντας το πλαίσιο Unsloth, δημιουργήσαμε 28.000 ομιλίες, και τις αξιολογήσαμε χρησιμοποιώντας το πλαίσιό μας, συγκρίνοντας τα βασικά και τα fine-tuned μοντέλα. Για τη στατιστική ανάλυση των αποτελεσμάτων εφαρμόσαμε t-tests και ANOVA tests. Τα αποτελέσματα δείχνουν ότι το fine-tuning παράγει στατιστικά σημαντικές βελτιώσεις στην πλειονότητα των μετρήσεων και οι νέες μας μετρήσεις καταδεικνύουν ισχυρή διακριτική δύναμη για τις πολιτικές διαστάσεις. Τμήματα της έρευνας που παρουσιάζονται σε αυτή τη διατριβή βασίζονται στην ακόλουθη δημοσίευση: "ParliaBench: Ένα Πλαίσιο Αξιολόγησης και Benchmarking για Κοινοβουλευτικούς Λόγους που Παράγονται από LLM" [20]. Αυτή η δημοσίευση συγγράφηκε σε συνεργασία με τον κ.Μάριο Κόνιαρη και τον κ.Παναγιώτη Τσανάκα και αναπτύχθηκε ως μέρος της έρευνας που διεξήχθη για αυτή τη διατριβή. Συνεισέφερα στην εννοιολόγηση, τη μεθοδολογία, την ανάλυση δεδομένων και την προετοιμασία του χειρογράφου.

Λέξεις-κλειδιά:

Παραγωγή Κοινοβουλευτικών Λόγων, Μεγάλα Γλωσσικά Μοντέλα, Επεξεργασία Φυσικής Γλώσσας, Παραγωγή Φυσικής Γλώσσας, Αξιολόγηση Μεγάλων Γλωσσικών Μοντέλων, Μεγάλα Γλωσσικά Μοντέλα ως Κριτές, Πολιτική Αυθεντικότητα, Ιδεολογική Ευθυγράμμιση, Αξιολόγηση Πλαισίου.

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 a benchmark for parliamentary speech generation. We constructed and preprocessed a dataset of speeches from UK Parliament Parlamin GB to enable systematic model training. We introduce a comprehensive 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. For linguistic quality and semantic coherence, we employed metrics including Perplexity, Self-BLEU, BERTScore, GRUEN Score, MOVER Score, and Distinct-n. We propose two novel embedding-based metrics, Political Spectrum Alignment and Party Alignment, to quantify ideological positioning. Additionally, we utilized the LLM-as-a-judge approach to evaluate six dimensions: conciseness, coherence, authenticity, political appropriateness, overall quality, and relevance. We fine-tuned five large language models (Mistral, Gemma, Qwen, Llama, Yi) using the Unsloth framework, for parameter-efficient training, generated around 28,000 speeches with the same context for each model, and evaluated them using our framework, comparing baseline and fine-tuned models. For statistical analysis of results, we applied t-tests and ANOVA tests. 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. Portions of the research presented in this thesis are based on the following publication: "ParliaBench: An Evaluation and Benchmarking Framework for LLM-Generated Parliamentary Speech" [20]. This paper was co-authored with Marios Koniaris and Panayiotis Tsanakas and developed as part of the work conducted for this thesis. I contributed to the conceptualization, methodology, data analysis, and manuscript preparation.

Keywords

Parliamentary Speech Generation, LLM Evaluation, Political Authenticity, Benchmark Evaluation, Natural Language Generation, Natural Language Processing, Ideological Alignment, Embedding-based Metrics, Parameter-Efficient Training.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Panayiotis Tsanakas, Dean of the ECE Department, who made this thesis possible. I am truly honored to have worked under his supervision.

I am equally grateful to my co-supervisor, Marios Koniaris, for his dedicated support and hands-on guidance throughout this research journey. His commitment to our weekly meetings, constructive feedback, and collaborative approach significantly enhanced the quality of this work. I am particularly thankful for his mentorship in co-authoring our research paper, which provided invaluable experience in academic writing and publication.

To my university friends who have supported me throughout my academic years, I am deeply thankful for your friendship, encouragement, and the countless moments of solidarity during challenging times. Your companionship made this journey far more meaningful and enjoyable.

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

Athens, October 2025

Argyro Tsipi

Table of Contents

Abstract	5
Abstract	7
Acknowledgements	9
1 Introduction	22
1.1 Evaluation Frameworks for LLM Political Speech Generation	22
1.2 Thesis Focus	22
1.2.1 Contributions	22
1.3 Chapter Organization	23
2 Related Work	24
2.1 Large Language Models (LLMs)	24
2.2 Speech Generation	25
2.3 Evaluation Framework	26
3 Theoretical Background	31
3.1 Dataset Specifics	31
3.2 Text Classification Techniques	32
3.3 Language Models	32
3.4 Finetuning	33
3.5 Speech Generation	36
3.6 Model Evaluation	36
4 Data Processing	38
4.1 Data Selection, Cleaning and Preparation	38
4.2 Data Pre-processing Pipeline	40
4.2.1 XML Parsing and Metadata Extraction	41
4.2.2 Speech-Party Temporal Alignment	42
4.2.3 Prompt Extraction	43
4.3 Political Orientation Statistics	44
4.4 Topic Categorization	44
4.5 Thematic Categorization	44
4.5.1 Thematic Classification Methodology	45
4.5.2 Thematic Classification Implementation	46
4.5.3 EuroVoc Categories	46
4.5.4 Topic Assignment Strategy	47
4.6 Statistics Plots	48

4.7	Training Data Format	48
4.7.1	Train-Test Split Implementation	49
5	Model Selection	50
5.1	Models	50
5.2	Mistral 7B v0.3 Model	50
5.2.1	Architecture Overview	50
5.2.2	Technical Specifications	51
5.3	LLaMA 3.1 8B Model	51
5.3.1	Architecture Overview	51
5.3.2	Technical Specifications	51
5.4	Gemma 2 9B Model	52
5.4.1	Architecture Overview	52
5.4.2	Technical Specifications	52
5.5	Yi 1.5 6B Model	53
5.5.1	Architecture Overview	53
5.6	Qwen2 7B Model	53
5.6.1	Architecture Overview	53
5.7	Model Selection Explanation	54
5.8	Tools and Environment	54
6	Fine-tuning	55
6.1	Unsloth Framework	55
6.2	Parameter-Efficient Fine-tuning with QLoRA	55
6.3	Trainer Implementation and Configuration	55
6.3.1	Multi-Model Chat Template Architecture	56
6.3.2	Training Arguments and Hyperparameter Selection	57
6.3.3	Training Pipeline	57
7	Speech Generation	59
7.1	Generation Pipeline Architecture	59
7.2	Input Mapping	59
7.3	Model Loading and Inference Configuration	60
7.4	Generation Parameters and Decoding Strategy	60
7.5	Multimodal Chat Template	61
7.6	Speech Validation	62
7.7	Examples of Generated Speeches	63
8	Evaluation Framework	68
8.1	Evaluation Methodology Overview	68
8.2	Linguistic Quality and Diversity Metrics	68
8.2.1	Perplexity	68
8.2.2	Distinct-N	68
8.2.3	Self-BLEU	69
8.3	Semantic Coherence and Text Quality Evaluation	69
8.3.1	GRUEN Score	69
8.3.2	BERTScore	71

8.3.3 MoverScore	71
8.4 Political Spectrum and Party Alignment	72
8.5 LLM-as-a-Judge Evaluation	73
8.5.1 Metric 1: Coherence	73
8.5.2 Metric 2: Conciseness	75
8.5.3 Metric 3: Relevance	76
8.5.4 Metric 4: Authenticity	77
8.5.5 Metric 5: Political Appropriateness	78
8.5.6 Metric 6: Overall Quality	80
8.5.7 Implementation Details	81
9 Statistical Methodology and Results	82
9.1 Performance Analysis	82
9.2 Statistical Significance Testing	82
9.3 Context Stability Analysis	83
9.4 Validation Analyses	84
9.5 Context Difficulty Analysis	84
9.6 Results	85
9.6.1 Overview and Fine-Tuning Impact	85
9.7 Political Context Results	93
9.7.1 Party alignment Patterns	93
9.7.2 Party Alignment Difficulty Analysis	94
9.8 Topic Results	95
9.8.1 Topic Performance Patterns	95
9.8.2 Topic Difficulty Analysis	95
9.9 Political Orientations Results	97
9.9.1 Political Orientation Patterns	97
9.9.2 Political Orientation Difficulty	98
10 Discussion	101
10.1 Summary of Key Findings	101
10.1.1 Architectural Design and Context Window Effects	101
10.1.2 Domain-Specific Fine-Tuning Impact	101
10.1.3 Novel Political Authenticity Metrics Validation	102
10.2 Limitations	102
10.2.1 Dataset and Scope Limitations	102
10.3 Future Research Directions	102
10.3.1 Near-Term Extensions	102
10.3.2 Multilingual and Cross-National Research	103
10.3.3 Real-Time Parliamentary Monitoring Systems	103
10.3.4 Bias Detection and Mitigation	104
10.3.5 Human-AI Collaborative Systems	104
10.4 Concluding Remarks	105
Bibliography	111

List of Figures

4.1	Overview of the ParlaMint-GB Preprocessing and EuroVoc Topic Classification Pipeline.	40
4.2	Party Changes Over Time	42
4.3	Summary Statistics	48
9.1	Finetuning Improvement by Method	87
9.2	Finetuning Impact	88
9.3	Absolute performance changes (fine-tuned vs. baseline) across evaluation categories. All metrics normalized to comparable scales.	89
9.4	Multidimensional Stability	90
9.5	Cross Context Stability	90
9.6	Overall performance comparison	91
9.7	Model comparison across each metric separately	91
9.8	LLM as a Judge correlation among metrics	92
9.9	Length Correlation Analysis	93
9.10	Party alignment scores (0-1 scale) for finetuned models across UK parliamentary parties. Color intensity indicates performance level	94
9.11	Party alignment difficulty scores (0-1 scale) for finetuned models across UK parliamentary parties. Dark green = most difficult	94
9.12	Model performance scores (0-1 scale) for finetuned models across topic domains. Color intensity indicates performance level	95
9.13	Topic difficulty scores (0-1 scale) for fine-tuned models across topic domains.	96
9.14	Topic Difficulty vs Consistency	97
9.15	Model performance scores (0-1 scale) for finetuned models across political orientations. Color intensity indicates performance level	98
9.16	Political orientation difficulty	99
9.17	Political Party and Topic performance by model	100

List of Tables

4.1	Cleaned ParlaMint-GB Dataset Statistics	39
4.2	Distribution of Party by Parliamentary House	39
4.3	Political Parties Represented in Dataset	44
4.4	Political Orientation Categories	44
4.5	Mapping between CAP Categories and EuroVoc Categories	45
5.1	Model Technical Specifications and Quantization Details	50
5.2	Comparison of Mistral 7B Model: Original and Optimized	51
5.3	Comparison of Llama 3.1 8B Model: Original and Optimized	52
5.4	Comparison of Gemma 2 9B Model: Original and Optimized	53
6.1	Complete QLoRA Configuration Parameters	55
6.2	Complete Training Configuration for Parliamentary Speech Generation	57
7.1	Party Distribution and Political Orientation	59
7.2	House Distribution in the Dataset	60
7.3	Model Generation Configuration Parameters	60
7.4	Baseline Model Speech - Gemma	63
7.5	Baseline Model Speech - Llama	63
7.6	Baseline Model Speech - Mistral	64
7.7	Baseline Model Speech - Qwen	64
7.8	Baseline Model Speech - Yi	64
7.9	Finetuned Model Speech - Gemma	65
7.10	Finetuned Model Speech - Llama	66
7.11	Finetuned Model Speech - Mistral	66
7.12	Finetuned Model Speech - Qwen	66
7.13	Finetuned Model Speech - Yi	67
9.1	Complete Performance Comparison between Baseline (B) and Fine-Tuned (F) Models by Conceptual Assessment Categories	85
9.2	Significant Pairwise T-Test Results (Bonferroni Corrected, $\alpha=0.05$)	86

Περίληψη

Η παραγωγή κοινοβουλευτικών λόγων παρουσιάζει συγκεκριμένες προκλήσεις για τα μεγάλα γλωσσικά μοντέλα πέρα από τις τυπικές εργασίες παραγωγής κειμένου. Σε αντίθεση με τη γενική παραγωγή κειμένου, οι κοινοβουλευτικές ομιλίες απαιτούν όχι μόνο γλωσσική ποιότητα αλλά και πολιτική αυθεντικότητα και ιδεολογική συνέπεια. Τα τρέχοντα γλωσσικά μοντέλα δεν διαθέτουν εξειδικευμένη εκπαίδευση για κοινοβουλευτικά πλαίσια και οι υπάρχουσες μέθοδοι αξιολόγησης επικεντρώνονται σε τυπικές μετρήσεις NLP παρά στην πολιτική αυθεντικότητα. Για να αντιμετωπίσουμε αυτό το πρόβλημα, παρουσιάζουμε ένα evaluation framework για την παραγωγή κοινοβουλευτικών λόγων. Κατασκευάσαμε ένα σύνολο δεδομένων ομιλιών από το Κοινοβούλιο του Ηνωμένου Βασιλείου για τη συστηματική εκπαίδευση μοντέλων. Εισάγουμε ένα πλαίσιο αξιολόγησης που συνδυάζει υπολογιστικές μετρήσεις με αξιολογήσεις LLM-as-a-judge για τη μέτρηση της ποιότητας παραγωγής σε τρεις διαστάσεις: γλωσσική ποιότητα, σημασιολογική συνοχή και πολιτική αυθεντικότητα. Προσθέτουμε επίσης δύο μετρήσεις, Party Align και Political Spectrum Analysis, για να ποσοτικοποιήσουμε την ιδεολογική τοποθέτηση. Εκπαιδεύσαμε με fine-tuning πέντε μεγάλα γλωσσικά μοντέλα (LLM), δημιουργήσαμε 28.000 ομιλίες και τις αξιολογήσαμε χρησιμοποιώντας το πλαίσιο μας, συγκρίνοντας τα βασικά και τα finetuned μοντέλα. Τα αποτελέσματα δείχνουν ότι το fine-tuning παράγει στατιστικά σημαντικές βελτιώσεις στην πλειονότητα των μετρήσεων και οι νέες μας μετρήσεις καταδεικνύουν ισχυρή διακριτική δύναμη για τις πολιτικές διαστάσεις.

Επεξεργασία Δεδομένων και Προετοιμασία Χρησιμοποιήσαμε το σύνολο δεδομένων ParlaMint-GB έκδοση 5.0 από το CLARIN, το οποίο περιλαμβάνει πρακτικά του Βρετανικού Κοινοβουλίου από τις 5 Ιανουαρίου 2015 έως τις 21 Ιουλίου 2022. Η επιλογή του συγκεκριμένου συνόλου δεδομένων βασίστηκε στο γεγονός ότι είναι γραμμένο στην αγγλική γλώσσα, αποφεύγοντας έτσι τις σημασιολογικές απώλειες που προκύπτουν από μεταφράσεις. Το σύνολο δεδομένων περιέχει δομημένα κοινοβουλευτικά πρακτικά με πλούσια μεταδεδομένα, συμπεριλαμβανομένων των ονομάτων των ομιλητών, των ρόλων τους, των πολιτικών τους τοποθετήσεων, των πληροφοριών φύλου, καθώς και πλήρη αντίγραφα των ομιλιών με τις αντίστοιχες ημερομηνίες τους. Το αρχικό σύνολο δεδομένων περιείχε 670.912 συνεισφορές από 1.951 μέλη του κοινοβουλίου σε 2.209 συνεδριάσεις, με συνολικό μέγεθος περίπου 135 εκατομμύρια λέξεις. Η διαδικασία καθαρισμού και προετοιμασίας των δεδομένων περιελάμβανε πολλαπλά στάδια φιλτραρίσματος. Διατηρήθηκαν μόνο τα πολιτικά κόμματα με περισσότερες από 1.000 ομιλίες, ενώ αφαιρέθηκαν ομιλίες με λιγότερες από 35 λέξεις (5ο εκατοστημόριο) και περισσότερες από 1.580 λέξεις (99ο εκατοστημόριο). Επίσης, αφαιρέθηκαν ομιλίες που σχετίζονταν με διαδικαστικά θέματα όπως "Business of the House" και "Point of Order", καθώς δεν προσέφεραν ουσιαστική πολιτική αξία. Μετά από αυτές τις διαδικασίες, το τελικό σύνολο δεδομένων περιείχε 447.778 ομιλίες από 1.901 μοναδικούς ομιλητές, που αντιπροσώπευαν 11 πολιτικές ομάδες, με συνολικό αριθμό περίπου 99,94 εκατομμύρια λέξεις. Ένα καινοτόμο στοιχείο της έρευνας ήταν η θεματική κατηγοριοποίηση των ομιλιών χρησιμοποιώντας το εργαλείο Kevlar για την αυτόματη ανάθεση κατηγοριών EuroVoc. Το Kevlar ταξινόμησε κάθε ομιλία σε μία από τις 21 θεματικές κατηγορίες που προέρχονται από τον θησαυρό EUROVOC, όπως Διεθνείς Σχέσεις, Νόμος, Κοινωνικά Ζητήματα, Πολιτική, Εκπαίδευση και Επικοινωνίες, Οικονομία, Ευρωπαϊκή Ένωση, Μεταφορές, Εμπόριο, Περιβάλλον, Ενέργεια και άλλες. Η επιλογή του Kevlar βασίστηκε στην εκπαίδευσή του σε νομικά, κυβερνητικά και πολιτικά κείμενα, καθώς και στη συμβατότητά του με την ταξινόμια EUROVOC. Το σύστημα επέστρεφε πολλαπλές έννοιες EuroVoc για κάθε ομιλία με βαθμούς εμπιστοσύνης, και διαλέξαμε την τελική θεματική κατηγορία με βάση τον τομέα με την υψηλότερη βαθμολογία.

Επιλογή και Αρχιτεκτονική Μοντέλων Η έρευνα εξέτασε πέντε διαφορετικά μεγάλα γλωσσικά μοντέλα, το καθένα με μοναδικά αρχιτεκτονικά χαρακτηριστικά και δυνατότητες. Το Mistral 7B v0.3 αποτελεί ένα προηγμένο μοντέλο transformer με 7,25 δισεκατομμύρια παραμέτρους που χρησιμοποιεί

Grouped Query Attention (GQA), Sliding Window Attention (SWA) και Rotary Position Embedding (RoPE). Η αρχιτεκτονική SWA περιορίζει τον υπολογισμό της προσοχής σε ένα σταθερό παράθυρο, μειώνοντας την υπολογιστική πολυπλοκότητα από τετραγωνική σε γραμμική κλιμάκωση ως προς το μήκος της ακολουθίας. Το Llama 3.1 8B της Meta περιλαμβάνει 8,3 δισεκατομμύρια παραμέτρους και υποστηρίζει πολυγλωσσικές δυνατότητες με παράθυρο πλαισίου 128.000 tokens. Χρησιμοποιεί GQA με 8 key-value heads και ένα εκτεταμένο λεξιλόγιο 128.000 tokens για αποδοτικότερη επεξεργασία κειμένου. Το Gemma 2 9B της Google, με 9,24 δισεκατομμύρια παραμέτρους, χρησιμοποιεί μια αρχιτεκτονική decoder-only transformer με 42 επίπεδα και εναλλασσόμενη τοπική και παγκόσμια προσοχή. Το Qwen2 7B περιλαμβάνει καινοτομίες όπως Dual Chunk Attention για επεξεργασία εκτεταμένων πλαισίων και YARN για βελτιωμένη εξαγωγή σε μεγαλύτερα μήκη ακολουθίας. Τέλος, το Yi 1.5 6B αποτελεί ένα δίγλωσσο μοντέλο 6 δισεκατομμυρίων παραμέτρων που προεκπαιδεύτηκε σε 3 τρισεκατομμύρια tokens με εξειδίκευση στα αγγλικά και τα κινέζικα. Όλα τα μοντέλα χρησιμοποιήθηκαν στις 4-bit quantized εκδόσεις τους από το Unsloth, μειώνοντας δραστικά τις απαιτήσεις μνήμης (μείωση 58-62%) και επιταχύνοντας την ταχύτητα (2,2-2,4 φορές ταχύτερα) ενώ διατηρούσαν ελάχιστη υποβάθμιση στην ποιότητα του μοντέλου.

Στρατηγική Fine-Tuning Η διαδικασία fine-tuning υλοποιήθηκε χρησιμοποιώντας το πλαίσιο Unsloth, ένα εργαλείο βελτιστοποίησης ειδικά σχεδιασμένο για αποδοτική προσαρμογή μεγάλων γλωσσικών μοντέλων. Εφαρμόστηκε η μέθοδος Quantized Low-Rank Adaptation (QLoRA), που ενσωματώνει quantization 4-bit με προσαρμογή πινάκων χαμηλής τάξης για αποδοτική προσαρμογή του μοντέλου. Η προσέγγιση αυτή επιλέχθηκε επειδή επιτρέπει την προσαρμογή μεγάλων προεκπαιδευμένων μοντέλων σε πολιτικές ομιλίες χωρίς να απαιτούνται τεράστιοι υπολογιστικοί πόροι ή να χάνονται οι γενικές γλωσσικές ικανότητες των μοντέλων. Η διαμόρφωση QLoRA περιελάμβανε LoRA rank (r) ίσο με 16, LoRA alpha ίσο με 16 για baseline κλιμάκωση, στόχευση 7 επιπέδων (όλες οι γραμμικές μετασχηματισμοί), μηδενικό LoRA dropout για ενεργοποίηση των βελτιστοποιήσεων Unsloth, και καμία διαμόρφωση bias για ταχύτερη εκπαίδευση και μειωμένη μνήμη. Τα μοντέλα εκπαιδεύτηκαν χρησιμοποιώντας Supervised Fine-Tuning σε ζεύγη οδηγιών-απαντήσεων, διαμορφωμένα χρησιμοποιώντας chat templates ειδικά για κάθε μοντέλο. Όλα τα μοντέλα εκπαιδεύτηκαν με ένα κοινό system prompt:

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

Κάθε μοντέλο χρησιμοποίησε τη δική του ειδική μορφή wrapping με ειδικά tokens και ακολουθίες ελέγχου. Το πεδίο instruction περιελάμβανε την ερώτηση prompt ή γενική οδηγία όταν δεν υπήρχαν διαθέσιμα prompts, ενώ το πεδίο context περιελάμβανε οργανωμένες πολιτικές πληροφορίες: πολιτικό κόμμα, θεματική κατηγορία EuroVoc, τμήμα κοινοβουλευτικής συζήτησης, πολιτική κατεύθυνση και Σώμα (Βουλή των Κοινοτήτων ή Βουλή των Λόρδων). Οι παράμετροι εκπαίδευσης περιελάμβαναν batch size 64, learning rate $2e-4$, μέγιστα βήματα 11.194 (2 epochs), warmup steps 336 (10% των μέγιστων βημάτων), AdamW optimizer, weight decay 0,01, μέγιστο μήκος ακολουθίας 1.024 tokens, και γραμμικό learning rate scheduler. Η εκπαίδευση διεξήχθη σε GPU instances A100 στο AWS.

Αρχιτεκτονική Συστήματος Παραγωγής Ομιλιών Το σύστημα παραγωγής ομιλιών σχεδιάστηκε να φορτώνει τα εκπαιδευμένα μοντέλα και να δημιουργεί πολιτικές ομιλίες βάσει δομημένων εισόδων. Για να εξασφαλιστεί συνέπεια και δικαιοσύνη στις συγκρίσεις μοντέλων, δημιουργήθηκε ένα ενοποιημένο σύνολο τμημάτων εισόδου και prompts από το test set, εγγυώμενο ότι όλα τα μοντέλα έλαβαν πανομοιότυπες ερωτήσεις και δεδομένα εισόδου. Διατηρήθηκαν ρεαλιστικά καταναμεητικά χαρακτηριστικά δειγματοληπώντας τα Houses σύμφωνα με την πραγματική τους αναπαράσταση στο Βρετανικό Κοινοβούλιο (78% Βουλή των Κοινοτήτων, 22% Βουλή των Λόρδων) και τα πολιτικά κόμματα σύμφωνα με τα αντίστοιχα βάρη

τους στο σύνολο δεδομένων (Συντηρητικοί: 59%, Εργατικό: 24%, Σκωτσέζικο Εθνικό Κόμμα: 5%, κ.λπ.). Αντιστοιχίσαμε τα sections με τα EuroVoc topics με τη διαδικασία Kevlar ώστε να διατηρηθεί μία ισορροπημένη και σχετική είσοδος για παραγωγή λόγου. Για την παραγωγή κειμένου, εφαρμόστηκε nucleus sampling (top-p) ως κύρια στρατηγική αποκωδικοποίησης. Αυτή η μέθοδος επιλέχθηκε έναντι παραδοσιακών μεθόδων όπως greedy search και beam search, καθώς αυτές συχνά παράγουν επαναλαμβανόμενο ή ασυνεπές κείμενο παρά την επίτευξη χαμηλού perplexity. Οι παράμετροι παραγωγής περιελάμβαναν 2.700 ομιλίες ανά μοντέλο, θερμοκρασία 0,7 για ισορροπία μεταξύ συνοχής και λεξιλογικής ποικιλίας, top-p 0,85 για εστιασμένες αλλά ποικίλες εξόδους, penalty επανάληψης 1,2, batch size 32, ελάχιστο αριθμό λέξεων 43, μέγιστο αριθμό λέξεων 635, και μέγιστα νέα tokens 850. Ως system prompt είχαμε το εξής:

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 parliamentary language appropriate for the specified House.
The speech should reflect the political orientation and typical positions of the specified party on the given topic.

Και σε αυτήν την περίπτωση, περάσαμε ως παραμέτρους εισόδου το prompt question, topic, section, party, house, political orientation.

Πλαίσιο Αξιολόγησης Η έρευνα υλοποίησε ένα ολοκληρωμένο πλαίσιο αξιολόγησης για την εκτίμηση της απόδοσης των μοντέλων σε πολλαπλές διαστάσεις. Η αξιολόγηση περιελάμβανε: (1) γλωσσική ποιότητα και ποικιλία, (2) σημασιολογική συνοχή και ποιότητα κειμένου, (3) πολιτική αυθεντικότητα, και (4) συνολική αξιολόγηση με LLM-as-a-Judge. Για τη γλωσσική ποιότητα, χρησιμοποιήθηκαν οι μετρήσεις Perplexity, Distinct-N και Self-BLEU. Το Perplexity μετρά πόσο «εκπλήσσεται» ένα γλωσσικό μοντέλο όταν συναντά συγκεκριμένο κείμενο, με χαμηλότερες βαθμολογίες να υποδεικνύουν φυσικότερο κείμενο. Χρησιμοποιήθηκε το GPT-2 base ως μοντέλο αξιολόγησης. Οι μετρήσεις Distinct-N αξιολογούν τη λεξιλογική ποικιλία μετρώντας την αναλογία μοναδικών n-grams προς τον συνολικό αριθμό tokens. Υπολογίστηκαν τα Distinct-1, Distinct-2, Distinct-3 και Distinct-4 χρησιμοποιώντας την τυπική διατύπωση σε επίπεδο λόγου. Το Self-BLEU μετρά την ομοιότητα μεταξύ παραγόμενων κειμένων από το ίδιο μοντέλο, με χαμηλότερες βαθμολογίες να υποδεικνύουν υψηλότερη ποικιλία. Η προσέγγιση λειτουργεί λαμβάνοντας κάθε παραγόμενη ομιλία και υπολογίζοντας τη βαθμολογία BLEU της σε σύγκριση με κάθε άλλη ομιλία στη συλλογή που δημιουργείται από το ίδιο μοντέλο με τις ίδιες εισόδους (party, topic, house, political orientation) και στη συνέχεια υπολογίζοντας τον μέσο όρο όλων αυτών των βαθμολογιών. Για τη σημασιολογική συνοχή, υλοποιήθηκαν τρεις κύριες μετρήσεις. Το GRUEN (Grammaticality, non-Redundancy, focUs, structurE, and coNherence) συνδυάζει grammaticality, non-redundancy (ως penalty) και focus (ως penalty), σε ένα ενοποιημένο μέτρο αξιολόγησης. Το BERTScore χρησιμοποιεί προεκπαιδευμένα BERT embeddings για να μετρήσει τη σημασιολογική ομοιότητα μεταξύ παραγόμενων και πραγματικών ομιλιών. Το MoverScore υπολογίζει τη σημασιολογική απόσταση μετρώντας το βέλτιστο κόστος μεταφοράς που απαιτείται για τη μετατροπή της κατανομής embeddings ενός κειμένου σε άλλο. Για την εύρεση των αναφορών των MoverScore, BertScore, για κάθε παραγόμενη ομιλία, επιλέξαμε τις N=6 πιο σημασιολογικά παρόμοιες ομιλίες από το αρχικό σύνολο δεδομένων ως αναφορές, μετρούμενες με βάση την ομοιότητα συνημίτονου μεταξύ των embeddings της κάθε ομιλίας του train dataset και των input prompts των παραγόμενων ομιλιών. Τα Political Spectrum Alignment (PSA) και Party Alignment είναι δύο μετρήσεις που υπολογίζονται χρησιμοποιώντας word embeddings. Το PSA μετρά πόσο καλά ευθυγραμμίζεται το παραγόμενο κείμενο με τον αναμενόμενο πολιτικό προσανατολισμό (σε μια κλίμακα 13 σημείων από

Left έως Right). Συγκεκριμένα, δημιουργούνται centroid embeddings από αυθεντικές κοινοβουλευτικές ομιλίες για κάθε πολιτικό προσανατολισμό, και στη συνέχεια το PSA score υπολογίζεται ως: $PSA = \text{cosine_similarity}(\text{speech}, \text{embedding}) \times \max(0, 100 - d/10 \times 100)$, όπου d είναι η απόλυτη απόσταση μεταξύ του αναμενόμενου και του προβλεπόμενου προσανατολισμού στην πολιτική κλίμακα, με τελικό εύρος τιμών 0-1. Το Party Alignment υπολογίζεται ως το cosine similarity μεταξύ του embedding της παραγόμενης ομιλίας και του party-specific centroid embedding, αξιολογώντας εάν το μοντέλο καταγράφει χαρακτηριστικά ειδικά για το κόμμα πέρα από τον γενικό ιδεολογικό προσανατολισμό, επίσης σε κλίμακα 0-1. Επιπλέον, χρησιμοποιήθηκε ένα σύστημα LLM-as-a-Judge με το μοντέλο Flow-Judge-v0.1 για την αξιολόγηση έξι διαστάσεων: coherence, conciseness, relevance, authenticity, political appropriateness και overall quality. Κάθε διάσταση αξιολογήθηκε σε κλίμακα 1-10 με λεπτομερή κριτήρια και rubrics.

Στατιστική Μεθοδολογία και Αποτελέσματα Η ανάλυση των αποτελεσμάτων περιελάμβανε εκτενείς στατιστικούς ελέγχους. Χρησιμοποιήθηκαν paired t-tests για συγκρίσεις μοντέλων στις ίδιες ομιλίες και independent t-tests για συγκρίσεις baseline έναντι fine-tuned μοντέλων. Υπολογίστηκαν μεγέθη επίδρασης (Cohen's d) και 95% διαστήματα εμπιστοσύνης για όλες τις διαφορές μέσω όρων. Εφαρμόστηκε διόρθωση Bonferroni για πολλαπλές συγκρίσεις για τον έλεγχο του familywise error rate.

Τα αποτελέσματα αποκάλυψαν ότι το fine-tuning βελτίωσε σημαντικά την απόδοση σε 45 από τις 70 συγκρίσεις μετρήσεων με στατιστική σημαντικότητα. Το Yi επέτυχε τις ισχυρότερες βελτιώσεις σε όλες τις διαστάσεις, ενώ το Llama παρουσίασε συνεπείς κέρδη. Το Gemma2 και το Qwen2 εμφάνισαν trade-offs ποιότητας, με βελτιώσεις σε μία κατηγορία να συνοδεύονται από μειώσεις σε άλλες, υποδεικνύοντας αρχιτεκτονικές διαφορές στον τρόπο που τα μοντέλα ισορροπούν ανταγωνιστικούς στόχους κατά το fine-tuning. Η ανάλυση σταθερότητας σε διαφορετικά πλαίσια αποκάλυψε ότι τα fine-tuned μοντέλα διατήρησαν συνεπή απόδοση σε πολιτικά πλαίσια (σύνθετη σταθερότητα 91,4-96,2%). Το Mistral επέτυχε την υψηλότερη συνέπεια (96,2%) παρά τα trade-offs στην απόλυτη απόδοση, ενώ το Llama (95,1%) εξισορρόπησε ισχυρή απόδοση με σταθερότητα.

Η ανάλυση ανά κόμμα έδειξε ότι τα μεγάλα κόμματα (Conservative, Labour) πέτυχαν σταθερή απόδοση σε όλα τα μοντέλα, επωφελούμενα από σημαντικό όγκο δεδομένων εκπαίδευσης. Τα μικρότερα κόμματα εμφάνισαν μεγαλύτερη μεταβλητότητα. Το Mistral αντιμετώπισε δυσκολίες με ετερογενείς ομάδες (Non-Affiliated: 0.436), ενώ το Qwen διαπρέπει με ιδεολογικά συνεπείς μειοψηφίες (Bishops: 0.664). Το Yi επέδειξε καλή απόδοση σε όλα τα κόμματα (0.614-0.633). Οι δύο μετρήσεις πολιτικής αυθεντικότητας (PSA και Party Align) διακρίνουν επιτυχώς τις πολιτικές διαστάσεις στόχο τους. Το Party Align διακρίνει τα κόμματα ενώ το PSA διακρίνει τους προσανατολισμούς (και τα δύο $p < 0.001$). Η ανάλυσή μας αποκαλύπτει ότι η απόδοση του Party Align εξαρτάται κυρίως από την αφθονία δεδομένων και την ιδεολογική συνοχή παρά από το μέγεθος του κόμματος. Τα μοντέλα μαθαίνουν επιτυχώς τα γλωσσικά μοτίβα ειδικά για κάθε κόμμα όταν τα δεδομένα εκπαίδευσης παρέχουν σαφή στυλιστικά σήματα, υποδεικνύοντας ότι η στοχευμένη συλλογή δεδομένων για υποεκπροσωπούμενα κόμματα θα μπορούσε να βελτιώσει την κάλυψη.

Η θεματική ανάλυση αποκάλυψε ότι το θέμα Science πέτυχε τις χαμηλότερες βαθμολογίες (μέσος όρος 0,516), ενώ τα Economics (0,610) και European Union (0,606) έδειξαν την υψηλότερη απόδοση. Τα θέματα Science και Geography κατατάχθηκαν ως τα πιο δύσκολα, ενώ τα Finance, Business και Economics κατατάχθηκαν στις χαμηλότερες. Οι τεχνικοί και φυσικοί επιστημονικοί τομείς εμφανίζουν μεγαλύτερη διαφωνία μεταξύ μοντέλων σε σχέση με τα οικονομικά και πολιτικά θέματα, γεγονός που συνάδει με τη μεγαλύτερη εξειδίκευση στην ορολογία και τις ταχέως εξελισσόμενες έννοιες. Αντίθετα, οι οικονομικές και πολιτικές συζητήσεις χρησιμοποιούν πιο σταθερά εννοιολογικά πλαίσια ευθυγραμμισμένα με τις βασικές κοινοβουλευτικές λειτουργίες.

Η απόδοση σε πολιτικούς προσανατολισμούς έδειξε αναμενόμενα μοτίβα. Η κεντρο-δεξιά θέση κυρι-

άρχησε στο σύνολο δεδομένων ωστόσο δεν πέτυχε από τις υψηλότερες βαθμολογίες. Εμφανίστηκαν δυνατά σημεία ειδικά για κάθε μοντέλο, με το Gemma και το Qwen να επιτυγχάνουν τις υψηλότερες βαθμολογίες στις Δεξιές θέσεις, ενώ το Mistral έδειξε αδυναμία σε όλες τις πολιτικές θέσεις το οποίο υποδεικνύει αρχιτεκτονική αδυναμία.

Συζήτηση και Συμπεράσματα Τα αποτελέσματα υπογραμμίζουν αρκετές βασικές ιδέες. Πρώτον, ο αρχιτεκτονικός σχεδιασμός παίζει καθοριστικό ρόλο στην πολιτική αυθεντικότητα, με τα εκτεταμένα context windows να επιτρέπουν βελτίωση. Δεύτερον, το domain-specific fine-tuning αναδεικνύεται ως κρίσιμος παράγοντας απόδοσης, καθώς 45 από τις 70 συγκρίσεις μετρήσεων έδειξαν στατιστικά σημαντικές βελτιώσεις. Τρίτον, η εισαγωγή νέων μετρήσεων πολιτικής αυθεντικότητας (PSA, Party Align) παρέχει μια συμπληρωματική αξιολογητική διάσταση πέρα από τις συμβατικές μετρήσεις NLP, επικυρωμένη μέσω της ανταπόκρισης στο fine-tuning και των discriminative tests ($p < 0.001$). Συλλογικά, τα ευρήματα αυτά καταδεικνύουν ότι το fine-tuning όχι μόνο βελτιώνει τη γλωσσική ποιότητα και τη σημασιολογική συνοχή, αλλά επιτρέπει επίσης στα μοντέλα να αναπαράγουν καλύτερα τις ιδεολογικές και στυλιστικές αποχρώσεις που χαρακτηρίζουν τον πραγματικό κοινοβουλευτικό λόγο. Το προτεινόμενο πλαίσιο αξιολόγησης εδραιώνει μια βάση για την ποσοτική αξιολόγηση της αυθεντικότητας σε εργασίες παραγωγής κειμένου με πολιτικό προσανατολισμό.

Εφαρμογές και Μελλοντικές Κατευθύνσεις Οι μέθοδοι και τα ευρήματα που παρουσιάζονται εδώ έχουν πιθανές εφαρμογές σε υπολογιστική κοινωνική επιστήμη, ψηφιακές ανθρωπιστικές επιστήμες και έρευνα διακυβέρνησης ΑΙ. Οι προσομοιωμένες κοινοβουλευτικές συζητήσεις μπορούν να χρησιμοποιηθούν για μελέτες μεγάλης κλίμακας της δυναμικής του λόγου, του πλαισίου επιχειρημάτων και της ευθυγράμμισης πολιτικής σε πολιτικά κόμματα. Επιπλέον, τέτοια μοντέλα θα μπορούσαν να υποστηρίξουν την ανάπτυξη εκπαιδευτικών και αναλυτικών εργαλείων που εκπαιδεύουν φοιτητές, δημοσιογράφους ή υπεύθυνους χάραξης πολιτικής να ασχολούνται κριτικά με την πολιτική ρητορική. Μελλοντικές επεκτάσεις περιλαμβάνουν: (i) πολυγλωσσική αξιολόγηση σε ευρωπαϊκά κοινοβουλευτικά σώματα για την αξιολόγηση της διαγλωσσικής γενίκευσης, (ii) ενσωμάτωση πρωτοκόλλων ανθρώπινης κρίσης για εξωτερική επικύρωση των αυτοματοποιημένων μετρήσεων, και (iii) συστηματικές έρευνες για πολιτική προκατάληψη, ιδεολογική απόκλιση και συνέπεια προοπτικής σε αντίθετες απόψεις. Συμπερασματικά, η παρούσα έρευνα επιδεικνύει ότι τα μεγάλα γλωσσικά μοντέλα, όταν προσαρμόζονται κατάλληλα σε πολιτικά δεδομένα, μπορούν να παράγουν αυθεντικές κοινοβουλευτικές ομιλίες που αντικατοπτρίζουν ιδεολογικές θέσεις και στυλιστικά χαρακτηριστικά. Το ολοκληρωμένο πλαίσιο αξιολόγησης που αναπτύχθηκε παρέχει εργαλεία για την ποσοτικοποίηση τόσο της γλωσσικής ποιότητας όσο και της πολιτικής αυθεντικότητας, ανοίγοντας νέες δυνατότητες για έρευνα στην υπολογιστική πολιτική επιστήμη και τη δημιουργία συνθετικού πολιτικού λόγου.

Introduction

1.1 Evaluation Frameworks for LLM Political Speech Generation

Recent advances in natural language processing, particularly the emergence of Large Language Models (LLMs), have created new opportunities for developing computational simulations designed to accurately replicate human behavior. In particular, LLMs have found application in political science research, including political speech generation. However, these models inherit biases from their training data, including political, and linguistic biases that can impact the authenticity and fairness of generated political content. In a similar way, LLMs can exhibit preferences for certain political positions, raising concerns about their suitability for political discourse. Despite the growing interest in applying LLMs to political communication tasks, existing evaluation methods typically only focus on general linguistic quality metrics such as fluency and coherence, but fail to capture the political authenticity, ideological consistency, stylistic appropriateness, and contextual relevance that are essential for political discourse. This gap in evaluation methodology impedes progress on reliably assessing LLM performance in political speech generation.

1.2 Thesis Focus

This thesis aims to provide a comprehensive framework for evaluating the quality of political speeches different models generate. Current evaluation frameworks for speech generation work well for general tasks, but they cannot properly conclude whether generated political content sounds authentic or fits the context. Because political speech is complex, it requires evaluation methods that can measure multiple aspects at once. Therefore, in our thesis we suggest evaluating them based on linguistic quality, semantic coherence, style and ideology analysis, and political credibility. Portions of the research presented in this thesis are based on the following publication: "ParliaBench: An Evaluation and Benchmarking Framework for LLM-Generated Parliamentary Speech". This paper was co-authored with Marios Korianis and Panayiotis Tsanakas and developed as part of the work conducted for this thesis. I contributed to the conceptualization, methodology, data analysis, and manuscript preparation.

1.2.1 Contributions

1. We trained five different models (Mistral, Llama, Gemma, Qwen and Yi) using political speech and speaker data from the ParlaMint Great Britain dataset.
2. We fine-tuned the models in an instruction-response style, using the UNSLOTH framework to accelerate the process.

3. We generated political speeches using the nucleus sampling method and by providing the models instructions that include information such as prompt questions, topics, political orientation, house and political parties to represent.
4. We evaluated the generated speeches by calculating Perplexity, Distinct-N, GRUEN Score, Mover-Score, BertScore, Self-Bleu Score, by performing embedding based analysis for Party Align and Political Spectrum and using a Large Language Model (Llama 3.1) as a Judge.

1.3 Chapter Organization

Related work on Large Language Models, speech generation and their evaluation frameworks is presented in Chapter 2. Chapter 3 provides definitions and explanations on the dataset we used, common speech generation and fine-tuning techniques, as well as model evaluation methods. Chapter 4 describes our ParlaMint dataset pre-processing step, while Chapter 5 delves into our language model selection. Our fine-tuning methodology using Unsloth is depicted in Chapter 6. Chapter 7 presents nucleus sampling as our political speech generation method. Chapter 8 details the implementation of multiple evaluation metrics including perplexity, distinct-n, BERTScore, GRUEN, MoverScore, SELF-BLEU, Party Align and Political Spectrum analysis and LLM-as-a-judge approach. Chapter 9 presents the results and comparative analysis across models and metrics. Finally, Chapter 10 summarizes conclusions and discusses future improvements.

Related Work

2.1 Large Language Models (LLMs)

Large Language Models in Politics and Democracy: A Comprehensive Survey

LLMs are being used to classify congressional bills with up to 83% accuracy, analyze environmental policies, and create collaborative policy design frameworks. They can simulate realistic institutional behaviors and enable stakeholder participation in policy development, though they tend to favor perspectives from developed countries.

In addition, LLMs excel at analyzing political texts in multiple languages, classifying sentiment and ideology with accuracy comparable to human coders. However, they raise concerns about manipulation. Specifically, AI-generated political messages can be as persuasive as human content, and LLM-powered search systems increase confirmatory querying by 15-43%, potentially amplifying echo chambers and political polarization.

In the field of political analysis, LLMs are capable in simulating voter behavior and predicting election outcomes, sometimes outperforming traditional models. They can facilitate deliberative democracy by mediating discussions and helping groups find common ground. However, they exhibit favor biases toward Western, educated, democratic populations.

LLMs are being adapted for military planning through specialized versions like Defense Llama. They can simulate historical conflicts and improve crisis management. However, concerning behaviors emerge in wargame simulations, including tendencies toward nuclear escalation, unpredictable aggressive actions even in neutral scenarios.

In order to simulate human interactions and economic behaviors, LLMs are creating "social simulacra". They can model epidemic spread, resource-sharing dilemmas, and behavioral economics scenarios with human-like decision patterns. Frameworks like GOVSIM and Social-LLM enable large-scale social system simulations.

On legal tasks, LLMs demonstrate strong performance with GPT-4 passing the Uniform Bar Examination. However, they suffer from "legal hallucinations". They tend to generate outputs that deviate from legal facts. Specialized legal LLMs like SaullLM are being developed to address these issues, while experts emphasize the need for careful ethical guidelines when deploying LLMs for legal advice [3].

LoRA: Low-Rank Adaptation of Large Language Models

Traditional fine-tuning updates all parameters in pre-trained models, making it computationally expensive for large models like GPT-3 with 175B parameters. Existing solutions like adapter layers add inference latency, and prompt-tuning methods reduce usable sequence length and are difficult to optimize.

Low-Rank Adaptation (LoRA) is a parameter-efficient method for adapting large language models to specific tasks without changing the original model weights. Instead of fine-tuning all parameters, LoRA adds small trainable matrices alongside frozen pre-trained weights using the formula:

$$W_0 + \Delta W = W_0 + BA, \quad (2.1)$$

where W_0 remains frozen (original weights), $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with $r \ll \min(d, k)$ (rank is much smaller than original dimensions).

This dramatically reduces trainable parameters while achieving performance comparable to full fine-tuning across RoBERTa, DeBERTa, GPT-2, and GPT-3. Some benefits include no additional inference time by merging the matrices $W = W_0 + BA$, 3 times less GPU memory usage, and easy task switching by swapping LoRA modules (the A,B matrices). The method works because adaptation requires only low-rank updates ($r = 1-4$ often sufficient), and LoRA learns to amplify task-specific features rather than repeating existing patterns [14].

2.2 Speech Generation

Political Speech Generation using Traditional NLP Techniques

Early work in political speech generation employs statistical NLP methods to generate political speech based on the political party (Republican/Democrat) and an opinion vote (Yes/No). The study uses the Convote dataset containing 3,857 speech segments from 53 US Congressional debates in 2005. After pre-processing (removing single sentences, HTML tags, correcting punctuation), the data was divided into four classes.

The system implements a language model that uses 6-gram statistical models to predict word sequences based on the probability of seeing the sixth word given the previous five. It also implements a topic model which employs Justeson and Katz POS tag filters to extract meaningful 2-3 word terms that characterize each political class. Topics are ranked by significance scores showing how often terms appear in specific classes compared to the entire corpus.

The system generates speeches by firstly randomly selecting an opening 5-gram from the target class. Then it continuously evaluates current topics in the generated speech using topic coverage metrics. It then combines language model probabilities and topic model probabilities with a weighting factor. It penalizes repeated phrases to ensure variety, and repeats the entire process until an end-of-speech token is generated.

The study tested some alternative approaches, including Recurrent Neural Networks (RNNs) but abandoned due to poor grammatical results and time constraints, Latent Dirichlet Allocation (LDA) which produced mostly generic, non-meaningful topics, and lastly, a sentence-based approach which concatenated whole sentences from training data based on structural and textual similarity measures, but results were less promising.

The results were evaluated manually based on grammatical correctness, sentence transitions, speech structure and speech content. In addition to the manual evaluation, it deployed an automatic evaluation as well, on grammar assessment through POS tag matching, and content evaluation by comparing topic coverage with real speeches from the same class.

Based on the results, the n-gram language models with topic-based word selection successfully generate grammatically correct speeches with good sentence transitions, but struggles with the speech content. This is where our study comes in to evaluate political speech generation using Large Language

Models [18].

2.3 Evaluation Framework

A Survey on Evaluation of Large Language Models

This thesis provides an overview of what to evaluate, where to evaluate, and how to evaluate Large Language Models.

Natural Language Tasks

Sentiment Analysis and Text Classification: Binary/triple classification with typically high performance.

Natural Language Inference (NLI): Determining logical relationships between premises and hypotheses.

Semantic Understanding: Interpreting meaning beyond surface-level text.

Reasoning: Complex inference tasks requiring models to deduce answers from incomplete information.

Natural Language Generation: Summarization, dialogue, translation, and question answering.

Multilingual Tasks: Performance across non-English languages (often overlooked).

Factuality: Alignment with real-world truths and avoiding "hallucinations".

Robustness, Ethics, and Trustworthiness

Robustness: Stability under unexpected inputs and adversarial attacks.

Ethics and Bias: Detection of toxic language, hate speech, and demographic stereotypes.

Trustworthiness: Addressing hallucinations and factual inaccuracies.

Domain-Specific Applications

Social Science: Political ideology analysis, computational social science tasks, legal case judgment.

Natural Science and Engineering: Mathematics (strong in basic operations, weak in advanced functions), chemistry, code generation.

Medical Applications: Medical queries, examination performance (USMLE), diagnostic assistance.

Agent Applications: Tool usage, educational assessment, search/recommendation systems, personality testing.

General Task Benchmarks

Chatbot Arena and MT-Bench: Real-world chatbot evaluation through user voting and multi-turn dialogues.

HELM: Holistic evaluation across multiple aspects.

MMLU: Comprehensive multi-task assessment BIG-bench: 204 challenging tasks across diverse domains.

Specific Downstream Benchmarks

Medical: MultiMedQA for medical knowledge.

Legal: CUAD for contract understanding.

Reasoning: ARB for advanced reasoning tasks.

Safety: TRUSTGPT for ethical dimensions, SafetyBench for security.

Multi-modal Benchmarks

MME and MMBench: Evaluating vision-language models.

SEED-Bench: 19,000 annotated questions for multimodal understanding.

Automatic Evaluation

Accuracy: Exact Match, F1 score, ROUGE score Calibration: Expected Calibration Error (ECE), Area Under Curve (AUC).

Fairness: Demographic Parity Difference (DPD), Equalized Odds Difference (EOD).

Robustness: Attack Success Rate (ASR), Performance Drop Rate (PDR).

Human Evaluation

Essential for tasks where automatic metrics are insufficient.

Number of Evaluators: Adequate representation for statistical significance.

Evaluation Criteria based on 3H rule (Helpfulness, Honesty, Harmlessness):

- Accuracy: Factual correctness
- Relevance: Appropriateness to context
- Fluency: Smooth, grammatical flow
- Transparency: Clear decision-making process
- Safety: Avoiding harmful content
- Human Alignment: Consistency with human values

Evaluator Expertise: Domain knowledge and methodological training

Performance Patterns

LLMs excel at basic language tasks but struggle with complex reasoning, advanced mathematics, and domain-specific applications. Supervised fine-tuning often outperforms zero-shot approaches despite fewer parameters.

Evaluation Challenges

Current evaluation methods face limitations in multilingual performance, factual accuracy, and bias detection. Human evaluation provides more comprehensive feedback but suffers from variance and cultural differences. Future Directions: The field is moving toward more holistic evaluation frameworks that combine automatic metrics with human judgment, emphasizing safety, fairness, and real-world applicability over pure performance metrics [6].

Systematic Biases in LLM Simulations of Debates

The study simulated political debates between LLM agents representing Republican and Democrat perspectives on four controversial American topics: Gun Violence, Racism, Climate Change, and Illegal Immigration. Researchers used three models (Mistral 7B, Solar 10.7B, and GPT-3.5) and created 40 agents for each political party with automatically generated background stories.

Debates followed a round-robin format with agents continuously rating their attitudes on a 0-10 scale before and after each discussion cycle. A "Default" agent with no political context was included to reveal the base model's inherent biases.

Despite being assigned specific partisan identities, agents consistently gravitated toward their base LLM's inherent biases rather than maintaining their assigned political perspectives. When the Default agent showed strong bias toward one side, the opposing partisan agent would significantly compromise their position. Even when the Default agent was excluded from debates, partisan agents continued shifting their attitudes in line with the model's inherent biases.

In a deviation from established human behavior, when Republican agents debated other Republicans (or Democrats with Democrats), they adopted more moderate positions rather than intensifying their beliefs. This directly contradicts the well-documented "Echo Chamber" phenomenon where like-minded individuals typically reinforce and escalate their beliefs through interaction.

Researchers developed an automated fine-tuning technique using only LLM-generated data. They created 100 political questions, generated 2,000 responses from agents, and used this dataset to fine-tune the base models using QLoRA.

Fine-tuning successfully altered the LLM's political orientation, causing all agents to shift their

behavior accordingly. When models were fine-tuned toward Republican perspectives, both Republican and Democrat agents moved rightward, and vice versa.

While fine-tuning did change political biases more permanently, it reduced general performance and capabilities of the models [29].

Large Language Models in Political Science: A Comprehensive Framework

This paper presents Political-LLM, a framework for integrating Large Language Models into computational political science research, mainly in political science applications and computational methodologies.

The framework identifies several applications in political science research. In predictive and generative tasks, LLMs are being used for sentiment analysis and public opinion tracking using datasets like OpinionQA, which contains 1,489 queries, and PerSenT for analyzing sentiment toward political entities. Election prediction capabilities leverage extensive datasets spanning decades, including U.S. Senate data from 1976-2020 and House data from 1976-2022, to analyze voting patterns and predict electoral outcomes. Legislative analysis represents another major application, with LLMs processing datasets like BillSum, containing over 22,000 U.S. Congressional bills, for automatic summarization of complex legislative documents.

Beyond predictive tasks, LLMs excel at behavioral simulation, allowing researchers to model voter behavior, policy negotiations, and political discourse dynamics. A significant case study using the 2016 ANES dataset demonstrated that larger models like GPT-4o and Llama 3.1-70B produced more accurate voting simulations that aligned closely with ground truth baselines of 47.7%. The framework also emphasizes causal inference capabilities, where LLMs use counterfactual generation to understand policy impacts and employ Chain-of-Thought reasoning for complex political analysis, such as assessing immigration reform impacts through historical, demographic, social, and economic factors.

The computational methodologies section details dataset preparation strategies. Data collection approaches involve broad sourcing from political speeches, legislative records, news articles, and social media platforms. The OpinionQA development process exemplifies this approach, utilizing Pew Research's American Trends Panel surveys across diverse demographics to create datasets. Annotation methods range from manual expert labeling to semi-automated processes with algorithmic assistance, and fully automated labeling with quality checks. However, there is still the need to address bias and ensure balanced representation across political viewpoints.

Data augmentation techniques include paraphrasing and synthetic data generation to expand scarce, limited political datasets. LLMs can generate hypothetical scenarios, such as voter opinion surveys based on historical data. The framework provides three detailed examples of dataset preparation: developing datasets for LLM-based political debiasing, automated annotation for legislative interpretation, and generating synthetic political datasets for election prediction research.

The fine-tuning process represents a crucial component of the framework, demonstrated through legislative summarization using the BillSum dataset. This process involves three key steps: data preprocessing to create well-aligned input-output pairs while removing irrelevant legal jargon, implementing parameter-efficient techniques like LoRA (Low-Rank Adaptation) or prefix-tuning to reduce computational costs, and conducting training optimization using GPU clusters with gradient accumulation and mixed-precision training. Prompt engineering plays a vital role, with examples including instructions to "read the following U.S. Congressional bill text and provide a summary highlighting main objectives, intended outcomes, and significant amendments" or to "summarize this bill in no more than 5 sentences, focusing on primary goals and authorized actions, avoiding technical jargon."

The framework also covers inference techniques, starting with zero-shot learning that enables pre-trained LLMs to perform political tasks without additional training. This approach is particularly valuable for sentiment analysis during events like the 2024 U.S. Presidential Election, using prompts such as "analyze the sentiment of the following statement about the presidential election as Positive, Negative, or Neutral." Few-shot learning represents a middle ground, using 3-4 carefully chosen examples for tasks like fake news detection while balancing prompt length with example diversity for optimal performance.

Advanced inference techniques include Retrieval-Augmented Generation (RAG), which combines knowledge retrieval with text generation to provide real-time political information by accessing up-to-date polling data and legislative updates. Knowledge editing allows for dynamic modification of internal model knowledge without requiring full retraining, particularly useful for updating policy information. Self-Consistency Decoding improves reliability by generating multiple responses and selecting the most consistent output, reducing randomness effects in complex political discourse analysis.

The framework identifies numerous benchmark datasets across multiple domains essential for political science research. For sentiment analysis, key datasets include OpinionQA, PerSenT, GermEval-2017, and various multilingual social media datasets that enable cross-linguistic and cultural sentiment studies. Election prediction datasets encompass U.S. Senate and House returns spanning 1976-2022, ANES surveys, and precinct-level data containing over 10 million data points. Legislative datasets include BillSum, CaseLaw, Federal Register, and DEU III covering three decades of EU legislative decisions. Additional specialized datasets cover misinformation detection (PoliFact, SciNews) and conflict resolution (Non-State Actors in Armed Conflict dataset, Peace Negotiations in Civil Conflicts).

The case study testing four LLMs (GPT-4o, GPT-4o-mini, Llama 3.1-8B, Llama 3.1-70B) on 2016 election data provides crucial insights into model performance. In voting simulation tasks, larger models achieved better accuracy and alignment with ground truth data, while smaller models showed pronounced bias toward winning parties. Importantly, the inclusion of political features helped mitigate inherent biases in model outputs. For feature generation quality, GPT-4o and Llama 3.1-70B achieved impressive response rates of 99.8% and 99.6% respectively, with generated political ideology distributions closely matching original ANES features, while smaller models demonstrated limited generation capabilities with consistent bias patterns.

The framework identifies several critical challenges facing the integration of LLMs into political science research. Data scarcity remains a significant issue, as political science lacks the large-scale, domain-specific datasets common in other fields. Bias and fairness concerns are particularly acute given the risk of amplifying biases from training data in sensitive political contexts. Explainability represents another major challenge, as the lack of transparency in decision-making processes limits trust and utility in high-stakes political research. Hallucination risks, where models generate plausible but factually incorrect information, pose serious threats to the integrity of political analysis.

To address these challenges, the framework proposes several innovative solutions. Pipeline integration involves developing modularized systems that break complex tasks into manageable components and integrating RAG systems for dynamic, up-to-date information retrieval. Data enhancement strategies include developing high-quality, domain-specific datasets, implementing synthetic data generation using LLM-based methods, and establishing collaborative partnerships for improved data access and curation.

Bias mitigation approaches encompass knowledge editing techniques to reduce bias in model outputs, counterfactual data augmentation to ensure fairness across different perspectives, and the implementation of explainable AI methods for transparency in predictions. Reliability improvements involve

feature attribution techniques like SHAP and Integrated Gradients, uncertainty quantification methods to measure output reliability, and validation checkpoints to reduce hallucination risks through cross-referencing with established datasets and domain expertise.

The framework emphasizes the critical need for evaluation criteria that go beyond traditional NLP metrics. The proposed evaluation framework includes four dimensions: policy relevance which assesses how effectively models contextualize policy positions in texts, electoral impact which measures accuracy in predicting public opinion, legislative influence which evaluating the societal implications of policies, and fairness and bias metrics that ensure balanced representation of all perspectives.

Finally, the framework addresses the important goal of democratizing political knowledge by making political information accessible to broader audiences. This involves developing LLM-driven tools for simplifying complex political language, providing multilingual support and cultural adaptation for diverse communities, creating interactive QandA systems for legislative documents, and integrating Explainable AI and Ethical AI principles to ensure transparency and accountability in political knowledge systems [23].

Theoretical Background

3.1 Dataset Specifics

CLARIN CLARIN (Common Language Resources and Technology Infrastructure) is a European research infrastructure that provides access to digital language resources and tools for humanities and social science researchers. CLARIN follows TEI and ISO standards for text encoding, ensuring consistent formatting across different datasets, languages and research projects [12].

ParlaMint ParlaMint is a multilingual corpus of parliamentary debates from multiple European countries from 2015 until 2022. ParlaMint follows XML-based encoding standards using the Text Encoding Initiative (TEI) guidelines. Each national subcorpus includes metadata covering speaker information, political party affiliations, parliamentary sessions, and temporal markers [9]¹.

ParlaMint-GB ParlaMint-GB represents a component of the ParlaMint corpus, containing parliamentary proceedings from the UK Parliament written in English. This subcorpus includes debates from both the House of Commons and House of Lords, with structured metadata providing detailed information about speakers, their political party affiliations, constituencies, parliamentary roles, and temporal information about sessions and speeches.

Topic Classification Topic classification is the process of automatically assigning texts to predefined categories or themes based on their content. In the context of parliamentary speeches, this involves analyzing speech transcripts and categorizing them according to thematic domains such as healthcare, education, defense, or economic policy.

Kevlar Kevlar is a multilingual classification toolkit specifically optimized for processing Public Administration acts. The system is designed to automatically assign EUROVOC thesaurus topics to documents. Kevlar uses BERT and RoBERTa models to train the models for classification [5].

EUROVOC Thesaurus EUROVOC is a multilingual thesaurus maintained by the European Union for document classification and retrieval. The thesaurus organizes concepts into hierarchical categories covering areas such as politics, law, economics, social issues, and international relations, with translations available in all official EU languages. In parliamentary corpus analysis, EUROVOC topics provide

¹T. Erjavec, M. Kopp, N. Ljubešić, et al. Multilingual comparable corpora of parliamentary debates ParlaMint 4.1. <http://hdl.handle.net/11356/1912>, jun 2024. Corpus: 8,073,406 utterances, 1,231,036,093 words. Project: <https://www.clarin.eu/content/parlamint>

a standardized framework for categorizing speeches, making possible to compare political discourse across different countries, and time periods ².

3.2 Text Classification Techniques

Zero-shot Classification Zero-shot classification is a machine learning approach that assigns labels to text without requiring training on labeled examples for the specific classification task. This method uses pre-trained language models that have learned general language understanding from large corpora and can classify texts into categories they have never seen during training [21, 25].

K-Means Clustering K-means clustering is an unsupervised machine learning algorithm that partitions data points into k clusters by grouping similar items together. The algorithm works by initializing k cluster centers, then iteratively assigning each data point to the nearest center and updating cluster centers based on the mean of assigned points until convergence [15, 19, 11].

N-gram N-grams is a feature extraction selection method that converts text into numerical features by extracting sequences of N consecutive words, enabling machine learning algorithms to process and analyze textual data [7].

Term Frequency-Inverse Document Frequency (TF-IDF) TF-IDF is a statistical method for evaluating the importance of words in a document relative to their frequency across an entire corpus. The technique calculates how often a word appears in a specific document and how rare the word is across all documents. Words that appear frequently in a particular document but rarely across the corpus receive higher TF-IDF scores, indicating they are characteristic of that specific document [7].

3.3 Language Models

Pre-trained Language Models Pre-trained Language Models undergo initial training on massive collections of unlabeled textual data to acquire fundamental linguistic patterns and structures. Following this pre-training phase, these models receive specialized training on smaller, task-oriented datasets through a process known as fine-tuning [35, 24].

Large Language Models Large Language Models consist of billions of parameters and are trained on large volumes of textual data. These models are pre-trained on text corpora and then evaluated in order to align with human preferences and values [35, 24].

Hugging Face Transformers Hugging Face Transformers is an open-source library that provides easy access to pre-trained language models for natural language processing tasks. The library offers a unified API for loading, fine-tuning, and deploying transformer-based models such as BERT, GPT, RoBERTa and supports both PyTorch and TensorFlow frameworks [32] ³.

²Publications Office of the European Union. EuroVoc: Multilingual thesaurus of the European Union. <https://op.europa.eu/en/web/eu-vocabularies/eurovoc>, 2025

³Hugging Face. Transformers: State-of-the-art machine learning for pytorch, tensorflow, and jax, 2023. URL <https://huggingface.co/docs/transformers/index>

3.4 Finetuning

Fine-tuning Fundamentals

Fine-tuning is a process that uses pre-trained models as starting points and adapts them to perform well on a more specific task. This methodology involves an additional training phase using smaller, task-specific datasets to tune the model's existing knowledge base to particular applications [24].

Fine-Tuning Types

Unsupervised Fine-tuning Unsupervised fine-tuning operates without labeled training data, instead exposing the model to large amount of unlabeled text from target domains to enhance its language understanding. This proves beneficial for specialized fields where labeled data are scarce, however, it is less precise for specific tasks [24].

Supervised Fine-tuning Supervised Fine-tuning employs labeled datasets specific to the target tasks, which requires a collection of text samples paired with their corresponding category labels. This methodology requires plenty labeled data, which can be costly to acquire [24].

Instruction-based Fine-tuning Instruction-based fine-tuning utilizes natural language prompts and instructions to guide model behavior. This approach is used to develop specialized conversational agents and assistants. It doesn't require extensive labeled datasets, but it needs quality instructional prompts [24].

Fine-Tuning Strategies

Task-Specific Fine-Tuning Customize the model for specific applications like summarizing text, generating code, classifying content, or answering questions by training on specialized datasets for each task [24].

Domain-Specific Fine-Tuning Adapt the model to understand and produce content within particular fields such as healthcare, finance, or law, ensuring it learns domain-specific terminology and concepts [24].

Parameter-Efficient Fine-Tuning (PEFT) Use cost-effective approaches like LoRA, QLoRA, and adapter layers that achieve good results while only modifying a small portion of the model's parameters, reducing computational requirements [24].

Half Fine-Tuning (HFT) Maintain a balance between preserving original knowledge and acquiring new capabilities by updating only half of the model's parameters in each training cycle, preventing complete override of pre-learned information [24].

Optimization for Fine-Tuning

Gradient Descent Variants

Gradient Descent Gradient Descent aims to find the best parameters for a neural network by updating them in every iteration based on the negative gradient of the cost function. It is efficient for smaller datasets as it can be computationally expensive [24].

Stochastic Gradient Descent Stochastic Gradient Descent reduces computations by updating parameters using a single or small subsets of training examples at each iteration. This approach is effective for large-scale datasets and online learning scenarios. It also requires a smaller learning rate compared to the classic Gradient Descent [24].

Mini-batch Gradient Descent Mini-batch gradient descent splits the data into small batches, and for every batch, it updates the parameters with averaged gradients. It is efficient mostly for medium sized datasets [24].

Adaptive Learning Rate Methods

AdaGrad The Adaptive Gradient Algorithm introduces parameter-specific learning rate adaptation based on historical gradient accumulation. This method reduces the number of updates for frequent parameters and maintains higher rates for sparse ones. This makes AdaGrad more effective for sparse datasets [24].

RMSprop Root Mean Square Propagation maintains steady learning rates by maintaining a running average of squared gradients and, therefore addressing AdaGrad's diminishing learning rate problem [24].

AdaDelta AdaDelta extends the adaptive learning rate concept by eliminating manual learning rate specification. The algorithm uses moving windows of gradient updates to compute adaptive learning rates, addressing both the diminishing learning rate problem and the requirement for learning rate hyperparameter tuning [24].

Momentum-Based Optimization

Adam Optimizer Adaptive Moment Estimation computes adaptive learning rates for each parameter by maintaining exponential moving averages of both gradients and squared gradients. It incorporates bias correction mechanisms and achieves faster convergence than alternative methods. The algorithm's effectiveness has established it as a default choice for many practitioners, although it requires attention to regularization to prevent overfitting [24].

AdamW AdamW represents an enhancement to the Adam optimizer that addresses overfitting by providing weight decay regularization. Particularly, it implements L2 regularization directly in parameter updates rather than through loss function modification [24].

Low-Rank Adaptation (LoRA) Fundamentals Low-Rank Adaptation is a parameter-efficient fine-tuning approach that keeps the original pre-trained weights frozen while adding smaller trainable decomposition matrices that learn new tasks. Instead of updating millions of parameters, LoRA only trains these lightweight decomposition matrices, thus reducing computational costs. Swapping out LoRA matrices allows us to create multiple task-specific versions from a single base model [24, 14].

Mathematical Foundation LoRA modifies pre-trained weight matrices through additive low-rank decomposition. Instead of updating the full weight matrix W , LoRA computes:

$$\hat{W} = W + \frac{a}{r} \cdot B \times A \quad (3.1)$$

where matrix A has dimensions $[r \times d_{input}]$, matrix B has dimensions $[d_{output} \times r]$, r is the rank (bottleneck dimension), and a is the scaling factor [14].

Quantized Low-Rank Adaptation (QLoRA) Quantized Low-Rank Adaptation extends LoRA principles through quantization strategies that compress model parameters to 4-bit precision from 16-bit. This compression technique transforms traditional 32-bit parameter storage to 4-bit representations and therefore achieves memory reductions. It is noteworthy to mention that it maintaining performance comparable to full-precision methods [24, 8].

Target Modules and Model Architecture Applying LoRA to all linear layers has proven to be more effective and match the full finetuning results. The attention layers include q_proj, k_proj, v_proj, o_proj. Whiel the MLP Layers are gate_proj, up_proj, down_proj.

LoRA's adjustable parameters include rank r , lora alpha, lora dropout, bias, and gradient checkpointing [24]. The rank produces more reliable results when assigned a larger value, however it can be slower and use more energy. Its values can range from 8 to 128. The lora alpha is thought to be reliable when it is set as the same value as the rank, but for more aggressive learning it is recommended to double the rank. The lora dropout is a regularization method that prevents overfitting by randomly zeroing out a portion of LoRA activations during training steps. However, lora dropout may not effectively regulate short training runs. Unless there is overfitting, setting lora dropout to zero will enable Unsloth's optimization and accelerate training. The bias argument trains the bias terms in the linear layers by adding extra trainable parameters. It is suggested to leave as zero to make finetuning faster. The gradient checkpointing parameter can be set to false, true or unsloth. The unsloth option significantly reduces memory usage.

Fine-tuning Hyperparameters

Learning Rate The learning rate defines how much and fast the model's weights are adjusted during each training step [24]. Higher learning rates can change the weights faster but can cause training instability or failure to find optimal minima. Meanwhile, lower learning rates require more training but can be more stable and precise. Typical range 2×10^{-4} to 5×10^{-6}

Epochs Epochs represent the number of times the model processes the complete training dataset [24]. More epochs allow better learning but risk overfitting through memorization of training data. However, fewer epochs reduce training time and prevent overfitting but may result in an under-trained model that hasn't captured underlying patterns. For most instruction-based datasets, 1-3 epochs are recommended as training beyond this may cause overfitting.

Batch Size and Gradient Accumulation Batch-size is a subset of the training dataset based on which we update the weights of the model [24]. The effective batch size is calculated as:

$$\text{Effective Batch Size} = \text{batch_size} \times \text{gradient_accumulation_steps} \quad (3.2)$$

The recommended practice is to use smaller `batch_size` and increase `gradient_accumulation_steps`.

Fine-Tuning Frameworks

UNSLOTH UNSLOTH is an open-source framework that makes fine-tuning large language models faster and cheaper. The framework can train models faster while using less computer memory compared to standard methods. The framework optimizes how models calculate attention mechanisms and implements gradient checkpointing, which reduces memory usage. The framework works with models like Llama, Mistral, and Gemma, and integrates with fine-tuning methods like LoRA and QLoRA. These technical improvements allow researchers to fine-tune large models on regular consumer hardware ⁴
⁵.

3.5 Speech Generation

Neural text generation systems face the degeneration problem, where traditional decoding strategies produce poor-quality output despite using well-trained language models. Approaches like beam search lead to repetitive and generic text, while sampling from the full probability distribution results in incoherent output. To address this issue, researchers propose Nucleus Sampling (top-p sampling), which dynamically selects tokens from the smallest subset containing a predetermined percentage (p) of the probability mass. Through evaluation using perplexity, diversity metrics, and human assessment, they show that this approach produces text that more closely resembles human writing [13].

3.6 Model Evaluation

Perplexity Perplexity, whose logarithm corresponds to the familiar entropy measure, represents a more appropriate measure of equivalent choice in speech recognition systems. Traditional complexity measures including vocabulary size, and branching factors have been shown to be inadequate for assessing the complexity of finite state grammars in speech recognition. Perplexity better captures recognition difficulty, and this measure can be applied even to languages without obvious statistical descriptions by finding entropy-maximizing probability assignments for any finite-state grammar [16].

GRUEN GRUEN focuses on the quality of the generated text instead of content selection like other evaluation metrics do. This evaluation framework assesses grammaticality, non-redundancy, focus, structure, and coherence using a BERT-based model combined with syntactic, semantic, and contextual features. GRUEN is unsupervised, deterministic, and adaptable to various tasks, showing high correlation with human judgments [38].

BERTScore BERTScore focuses on meaning rather than just string matching. Traditional metrics like BLEU rely on n-gram overlap and fail to recognize which words account for the same meaning. BERTScore addresses these limitations by computing sentence similarity using contextualized token embeddings from pre-trained BERT models, calculating cosine similarities between tokens [34].

⁴nc. Unsloth. Fine-tuning pretrained language models (unsloth llms guide). <https://docs.unsloth.ai/get-started/fine-tuning-llms-guide>, 2025

⁵D. Han, M. Han, and Unsloth team. Unsloth: Fast, memory-efficient fine-tuning of LLMs. GitHub repository. <https://github.com/unslothai/unsloth>, 2023

MoverScore MoverScore addresses limitations in traditional text generation evaluation by measuring semantic similarity through optimal transport theory. Rather than relying on n-gram overlap like BLEU and ROUGE, MoverScore computes the minimum semantic transportation cost needed to align system-generated text with reference text using contextualized embeddings [36].

LLM as a Judge with MT Bench and Chatbot Arena The LLM-as-judge approach is introduced as an alternative to the computationally expensive human evaluation for AI chat assistants. There are three main approaches.

Pairwise comparison presents an LLM judge with a question and two responses, asking it to determine which is better or declare a tie.

Single answer grading involves directly scoring individual responses without comparison.

Reference-guided grading provides correct solutions alongside the responses being evaluated, particularly useful for math and factual questions.

LLM judges might provide scalability and explainability, however, they come with position, name, verbosity and self-enhancement bias. In position bias, judges prefer the first response presented. Some models showed name bias, favoring responses labeled "Assistant A" regardless of content quality. In verbosity bias, LLM judges favor longer responses even when shorter ones are more concise. Meanwhile, LLMs with self-enhancement bias favor responses generated by themselves. LLM judges also demonstrate limited capability in grading math and reasoning questions. Surprisingly, they struggle even with basic math problems they can solve independently because they are being misled by incorrect answers in the context.

However, position bias can be controlled by swapping response positions and requiring consistent judgments across both orderings, or by using random positioning at scale. For math and reasoning problems, chain-of-thought prompting asks judges to solve problems independently before grading, though this doesn't fully solve the issue since judges can still be misled by context. Reference-guided methods prove most effective, as they have the judges generate their own solutions first and then using these as references during evaluation. When biases are properly addressed, GPT-4 judges achieve over 80% agreement with human evaluators, which matches the level of agreement between different human evaluators [37].

Data Processing

4.1 Data Selection, Cleaning and Preparation

For this thesis, the ParlaMint-GB dataset version 5.0 from CLARIN was chosen as the foundation for our experiments [9]. While ParlaMint includes data from several national parliaments, we selected the UK corpus (ParlaMint-GB) specifically because it is natively in English, eliminating potential semantic loss that comes with using translated versions. Additionally, we opted to work with the raw version of the dataset rather than the linguistically annotated one to retain full control over pre-processing and maintain fidelity to the original content.

Our chosen dataset (version 5.0) consists of structured UK parliamentary proceedings, including metadata such as speaker names and roles, political affiliations, gender information, links to their respective social media pages where applicable, and complete speech transcripts along with their dates. It is worth noting that the speech transcripts also contain marked-up transcriber comments. This dataset is suitable for our political language modeling due to its structured, and XML formatted data [9].

The Parliament of the United Kingdom comprises two Houses, Commons and Lords. The corpus contains proceedings from the House of Commons and the House of Lords between 5th January 2015 to 21st July 2022. The corpus itself contains 670,912 contributions (utterances) from 1,951 members of parliament and peers across 2,209 meetings. The total size of the corpus is around 135 million words.

We only kept the political parties that have more than 1000 speeches. We removed speeches with less than 35 words (5th percentile), and over 1580 words (99th percentile). We also filtered out the "Unknown" party. We removed speeches with sections named "Business of the House" and "Point of Order", as they produced no essential value. From the speeches we replaced the left and right double quotation marks (U+201C and U+201D) with regular double quotation marks (U+0022).

Statistic	Value
Total number of speeches	447,778
Number of unique speakers	1,901
Number of political affiliations	11
Total words	~99.94 million
Mean words per speech	223.2
Median words per speech	99.0
Standard deviation	278.7
Minimum	36.0
Maximum	1,579.0
5th percentile	43.0
10th percentile	49.0
25th percentile	66.0
75th percentile	242.0
90th percentile	619.0
95th percentile	872.0
99th percentile	1,322.0

Table 4.1. *Cleaned ParlaMint-GB Dataset Statistics*

Party	Commons	Lords	Total
Conservative	223,712	39,801	263,513
Labour	85,661	23,170	108,831
Scottish National Party	23,562	0	23,562
Liberal Democrats	8,017	15,500	23,517
Crossbench	0	11,878	11,878
Democratic Unionist Party	6,228	382	6,610
Independent	2,369	414	2,783
Plaid Cymru	1,817	412	2,229
Green Party	782	1,210	1,992
Non-Affiliated	0	1,713	1,713
Bishops	0	1,150	1,150
Total	352,148	95,630	447,778

Table 4.2. *Distribution of Party by Parliamentary House*

4.2 Data Pre-processing Pipeline

The data pre-processing step converts the ParlaMint-GB XML files into formats ready for training.

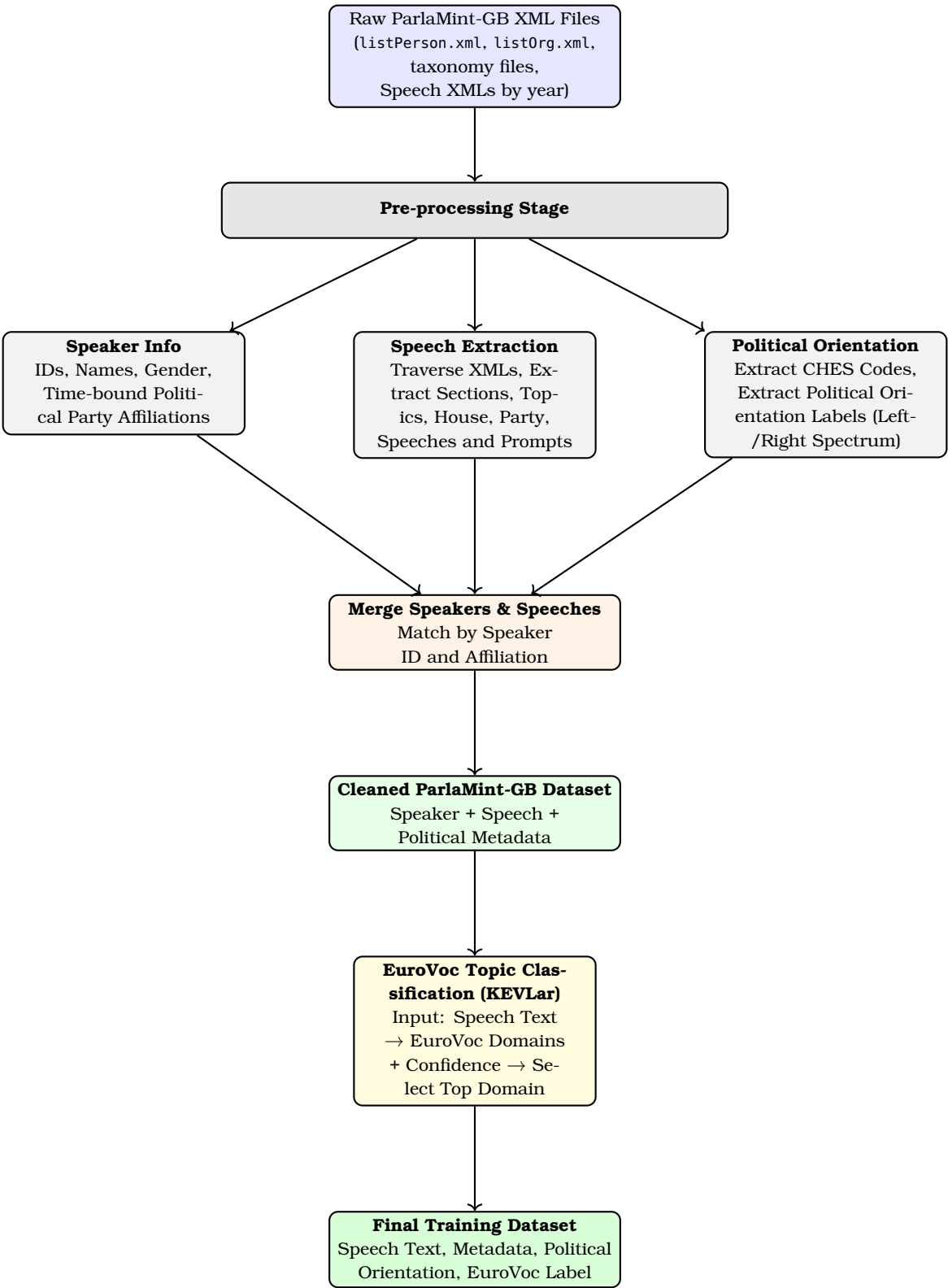


Figure 4.1. Overview of the ParlaMint-GB Preprocessing and EuroVoc Topic Classification Pipeline.

4.2.1 XML Parsing and Metadata Extraction

Initially, for our pre-processing stage we handle the XML structure and extract useful training data using Python. We made use of the metadata provided in the raw ParlaMint-GB dataset. Each speaker is represented in the `listPerson.xml` file, which includes key attributes such as full name, unique identifier, gender, and a list of political affiliations [9]. The speech content XML files were organized in folders named by year and contained speech metadata, such as date of the speech, the debate section, the CAP topic, the house and the speaker ID. We categorized the speeches in Parliaments based on whether the xml file ended on 'lords' for House of Lords or 'commons' for House of Commons. `ParlaMint-taxonomy-speaker_types.xml` contains the information on speaker types: Chairperson, Regular, Guest along with their respective descriptions. `ParlaMint-taxonomy-topics.xml` contains the 23 Comparative Agendas Project major topic labels that are used to categorize the speeches: Agriculture, Civil Rights, Culture, Defense, Domestic, Commerce, Education, Energy, Environment, Foreign Trade, Government Operations, Health, Housing, Immigration, International Affairs, Labor, Law and Crime, Macroeconomics, Mix, Other, Public Lands, Social Welfare, Technology, Transportation. `ParlaMint-listOrg.xml` contains the party names, their political orientations and their wikipedia pages. `ParlaMint-taxonomy-CHES.xml` contains the taxonomy of identifiers from the Chapel Hill Expert Survey (CHES) trend files: 1999-2019 Codebook. And finally, `ParlaMint-taxonomy-politicalOrientation.xml` contains the political orientation of political parties and parliamentary groups: Left, Centre, Right, Far-left, Far-right, Centre-left, Centre-right, Centre to centre-left, Centre to centre-right, Centre-left to left, Centre-right to right, Left to far-left, Right to far-right. The pre-processing pipeline extracted both speaker metadata and speech content through the following components:

Speaker Information Processing

- Parsed unique speaker identifiers and full names
- Extracted political affiliations with temporal bounds

Speech Content Extraction

- Traversed dated session XML files to extract individual speeches
- Extracted debate section topics and speech topic context information based on XML structure and `<head>` elements
- Filtered out procedural elements and non-substantive content

Political Orientation Extraction

- Extracted the political orientation code for each party from `ParlaMint-listOrg.xml`
- Matched the political orientation code to the political orientation label from `ParlaMint-taxonomy-politicalOrientation.xml`

We matched the speakers with their speeches based on the speaker ID attribute each speech had, along with their political orientations.

4.2.2 Speech-Party Temporal Alignment

To ensure that each speech was matched to the correct political party at the time of delivery, we implemented a temporal matching mechanism that handled political party changes and role transitions of speakers over time.

Alignment Algorithm

1. Extract the date of each speech from the XML session files
2. Parse each speaker's list of political affiliations from `listPerson.xml`
3. Extract the time validity range (`@from` and `@to` attributes) indicating when that affiliation was active
4. Compare the speech date with the affiliation's `from` and `to` dates to identify the active affiliation
5. In cases where no exact match was found (e.g., missing `to` date), default to the most recent affiliation that started before the speech date

This temporal alignment was important in order to accurately attribute party positions over time, especially for members who may have switched affiliations during their parliamentary careers.

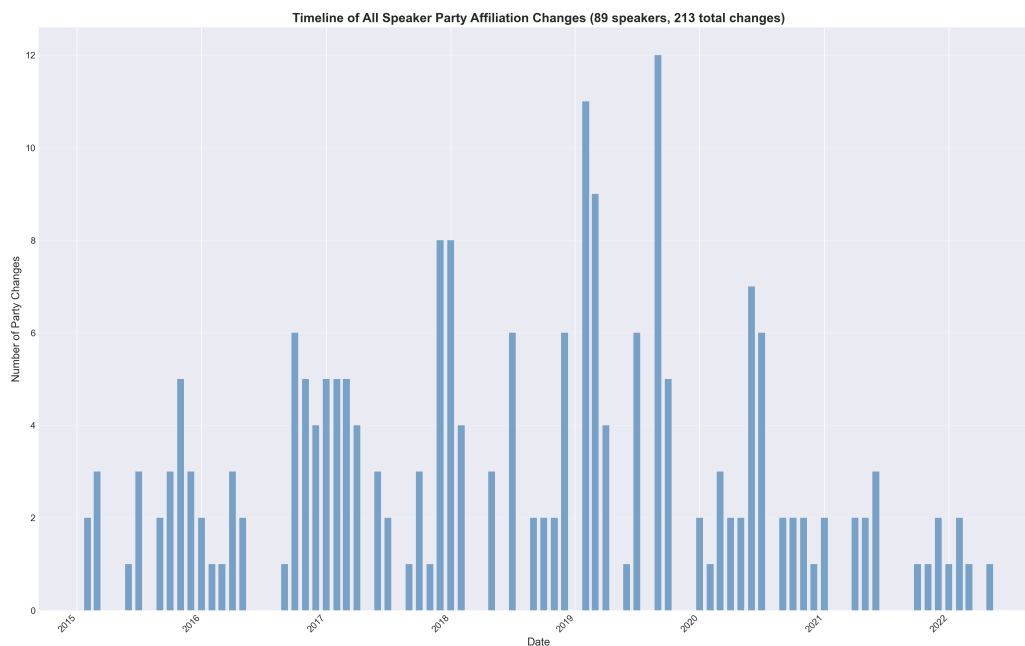


Figure 4.2. Party Changes Over Time

4.2.3 Prompt Extraction

The extracted speech content included question prompts and speeches. We identified patterns where question prompts began with a letter or number, allowing us to separate them from actual speeches. During extraction, we stored the speeches as string values in the "speech" attribute, while the individual question prompts were collected into a list of strings under the "prompts" attribute, as a single speaker could have expressed more than one question. We attributed the prompts to the speeches with the same debate section. Finally, we cleaned the prompts by removing the number and letter prefixes.

The following is a sample of the generated JSON file:

```
{
  "speaker": "Jennifer Jones",
  "speech": "My Lords, will the Minister consider some recommendations from a recent report from the Econo",
  "section": "Employment: Internships",
  "section_type": "debateSection",
  "speech_date": "2015-01-14",
  "speech_id": "ParlaMint-GB_2015-01-14-lords.u46",
  "filename": "ParlaMint-GB_2015-01-14-lords.xml",
  "party_affiliation": "party.GP",
  "party_name": "Green Party",
  "political_orientation": "L",
  "political_orientation_desc": "Left",
  "ches_lrgen": "2.0",
  "party_affiliation_from": "2013-09-20",
  "party_affiliation_to": "9999-12-31",
  "party_role": "member",
  "house": "House of Lords",
  "parliament_affiliation_from": "2013-09-20",
  "parliament_affiliation_to": "9999-12-31",
  "parliament_role": "member",
  "is_government_member": false,
  "government_role": null,
  "government_affiliation_from": null,
  "government_affiliation_to": null,
  "speaker_type": "regular",
  "speaker_type_desc": "Regular Speaker",
  "speech_topic_codes": ["labor"],
  "speech_topic_descriptions": ["Labor"],
  "section_topic_codes": [],
  "section_topic_descriptions": [],
  "prompts": []
}
```

4.3 Political Orientation Statistics

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 4.3. *Political Parties Represented in Dataset*

Orientation	Parties Included	Total Speeches	Percentage
Centre-right	Conservative	263,513	58.85%
Centre-left	Labour, Scottish National Party	132,393	29.57%
Centre to centre-left	Liberal Democrats	23,517	5.25%
Unknown	Bishops, Crossbench, Independent, Non-Affiliated	17,524	3.91%
Right	Democratic Unionist Party	6,610	1.48%
Centre-left to left	Plaid Cymru	2,229	0.50%
Left	Green Party	1,992	0.44%
Total		447,778	100.00%

Table 4.4. *Political Orientation Categories*

4.4 Topic Categorization

ParlaMint 5.0 version implemented two thematic categorizations, one speech-level and one section-level. The speech-level topics include the 23 categories mentioned in the Data Selection and Preparation subsection, and include one speech per topic. However, the section level topics are unique debate section topics that include more than one speech per debate section topic.

4.5 Thematic Categorization

While ParlaMint uses Comparative Agentas Project (CAP) classification¹, we selected EuroVoc² 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 Kevlar classification methodology. We argue that this approach is particularly well-suited for our dataset because it was specifically designed

¹Comparative Agendas Project. Comparative agendas project. <https://www.comparativeagendas.net>, 2025

²Publications Office of the European Union. EuroVoc: Multilingual thesaurus of the European Union. <https://op.europa.eu/en/web/eu-vocabularies/eurovoc>, 2025

for legal and governmental texts. For speeches yielding multiple concepts, we selected the highest individual concept score.

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 4.5. Mapping between CAP Categories and EuroVoc Categories

4.5.1 Thematic Classification Methodology

After evaluating several approaches, Kevlar, a multilingual text classification toolkit, was chosen to assign EuroVoc topics to speeches due to its training on legal, governmental, and policy-oriented texts, and compatibility with EUROVOC taxonomy [5].

Alternative Approaches Evaluated Zero-shot Classification: We initially explored a zero-shot classification approach using the pre-trained “facebook/bart-large-mnli” model from Hugging Face Transformers. This method involved loading a pre-trained classifier and defining a set of EUROVOC topic categories, then automatically assigning each parliamentary speech to the most semantically similar category without requiring any training data [21]. We tested this approach on a sample of 300 speeches from the House of Commons, with the classifier analyzing each speech’s content and selecting the appropriate topic label. While this supervised approach provided direct topic assignment, we decided against using it for our full dataset due to computational constraints.

Manual Cluster Mapping: We then tested a manual labeling approach in which we labeled speeches to train a classification model. In this method we created a sampling function that selected random speeches within each political party. The system showed each speech to a person who manually assigned it to one of 21 topic categories by selecting a number from a list. We labeled 5 speeches per political party and combined all the labeled data into one training dataset. We then used

these manually labeled data to train the "*distilbert-base-uncased*" model for automatic topic classification. In order to achieve this, we converted topic labels to numbers, split the data into training and testing sets, and trained the model for 4 epochs. Unfortunately, this approach produced poor results because our manually labeled data were very few compared to our large collection of parliamentary speeches, making them insufficient for the model to learn meaningful patterns, topics and speaking styles. Additionally, manually labeling speeches one by one would have taken far too long for our large dataset, making this method both ineffective and impractical for our research.

K-Means Clustering: We lastly explored an unsupervised K-means clustering approach for topic classification of parliamentary speeches. This method involved extracting and combining speech content with section metadata, then vectorizing the text using TF-IDF, which calculates the importance of each word in a speech. [19] We applied K-means clustering with eight clusters (topics), which we chose based on the top keywords, to automatically group speeches by thematic similarity. The algorithm successfully identified the following eight clusters: "Governance and Domestic Policy", "Parliamentary Procedure and QandA", "Northern Ireland and Brexit Protocol", "Trade and Brexit", "Legislation: EU Withdrawal and Referendum", "House of Lords and Amendments", "Health and Social Care", and "House Business and General Debate". We tried mapping and assigning these clusters to EuroVoc categories. However, the generated clusters, though coherent, did not directly correspond to the standardized EuroVoc taxonomy. Therefore, we decided against using K-means for our final classification system, even though it provided interpretable cluster keywords.

4.5.2 Thematic Classification Implementation

Kevlar was ultimately selected because of its ease of use, its computational efficiency compared to the alternative approaches we tried, and its built-in mappings to EuroVoc domain categories [5]. The implementation process involved:

Kevlar Integration Process

1. Prepare cleaned speech text for topic classification
2. Submit speech content to Kevlar's multilingual classification API
3. Process returned EuroVoc concept predictions with confidence scores

4.5.3 EuroVoc Categories

The following 21 thematic categories, derived from the EUROVOC thesaurus, were used for speech classification ³:

1. International Relations
2. Law
3. Social Questions
4. Politics
5. Education and Communications

³Publications Office of the European Union. EuroVoc: Multilingual thesaurus of the European Union. <https://op.europa.eu/en/web/eu-vocabularies/eurovoc>, 2025

6. Geography
7. Economics
8. Employment and Working Conditions
9. European Union
10. Transport
11. Trade
12. Environment
13. Production, Technology and Research
14. Energy
15. Agriculture, Forestry and Fisheries
16. Finance
17. Industry
18. Business and Competition
19. Agri-foodstuffs
20. International Organisations
21. Science

4.5.4 Topic Assignment Strategy

Kevlar returns multiple EuroVoc concepts per speech, each with a confidence score and associated domain label, which corresponds to a higher-level thematic category [5]. Since many concepts can be detected in a single speech, a rule was needed to determine a single, final topic label suitable for classification and training. We assigned the topic label corresponding to the domain with the highest individual score.

4.6 Statistics Plots

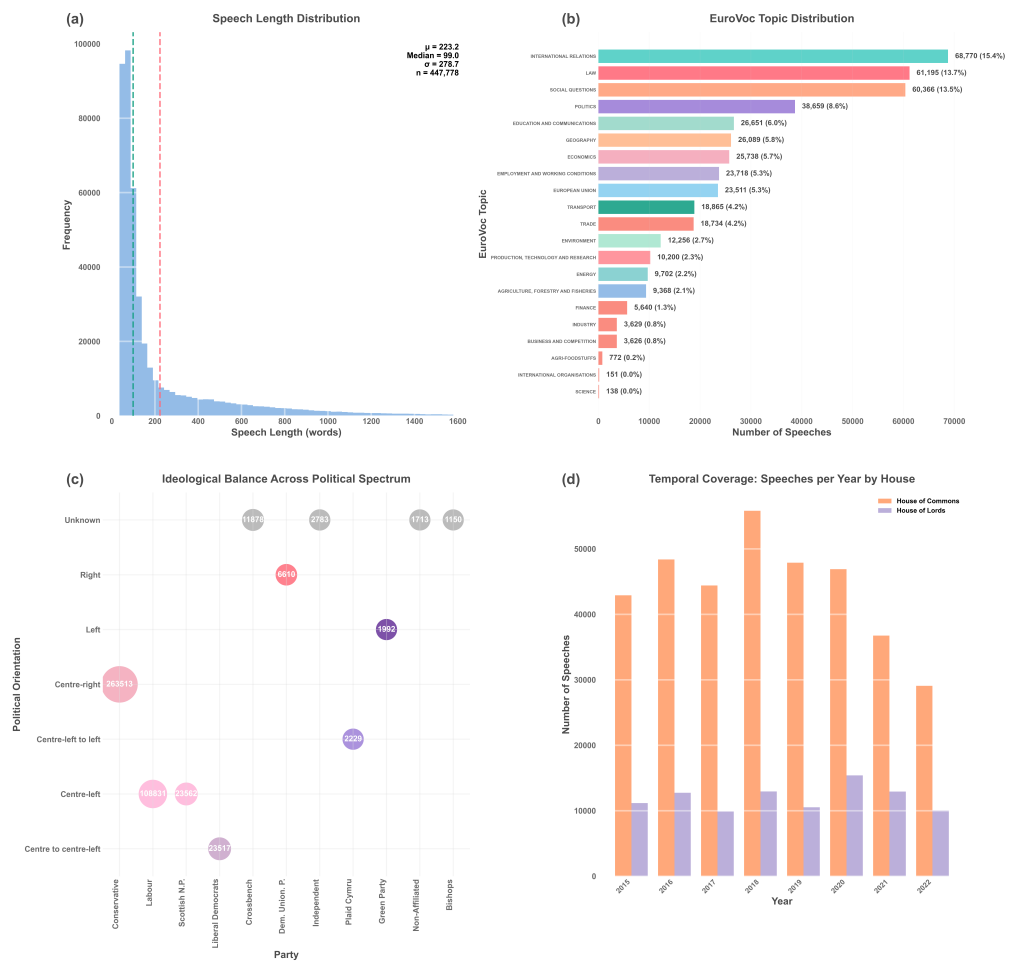


Figure 4.3. Summary Statistics

4.7 Training Data Format

The final input structure for training conversational models consisted of five components integrated into a JSONL (JSON Lines) format:

```
{
  "speech_1": {
    "speaker": "David Nuttall",
    "speech": "I have a couple of points to make on this group of amendments Amendment 7 deals with a requirement",
    "section": "Local Government (Religious etc. Observances) Bill",
    "speech_date": "2015-01-16",
    "speech_id": "ParlaMint-GB_2015-01-16-commons.u34",
    "filename": "ParlaMint-GB_2015-01-16-commons.xml",
    "party": "Conservative",
    "topic": ["Government Operations"],
    "prompts": [],
    "house": "House of Commons",
```

```
"political_orixentation_code": "CR",  
"political_orientation_label": "Centre-right",  
"eurovoc_topic": "POLITICS"  
},  
...  
}
```

However, for training our models, we passed the following parameters only: speech, section, party, prompts, house, political orientation label, EuroVoc topic. This is because each training instance was structured to enable the model to learn speech generation conditioned on both political party and topic context. Our goal was to simulate how representatives of specific parties would respond to given prompts by producing contextually aligned political discourse.

4.7.1 Train-Test Split Implementation

We implemented the following stratified train-test split:

- Training set: 80% of total records
- Test set: 20% of total records
- Random seed: 42 for reproducibility

Model Selection

5.1 Models

The goal was to explore how different large language models perform in political dialogue generation. We chose these models based on their different design architecture, performance scores, and compatibility with the UNSLOTH finetuning method.

Model	Quantized Version	Memory Reduction	Inference Speed
Mistral 7B v0.3	mistral-7b-v0.3-bnb-4bit	62%	2.2×
Llama 3.1 8B	Meta-Llama-3.1-8B-bnb-4bit	58%	2.4×
Gemma 2 9B	gemma-2-9b-bnb-4bit	58%	2.2×
Qwen2 7B	Qwen2-7B-bnb-4bit	N/A	N/A
Yi 1.5 6B	Yi-1.5-6b-bnb-4bit	N/A	N/A

Table 5.1. Model Technical Specifications and Quantization Details

Table 5.1 summarizes technical specifications and quantization details for the selected models.

5.2 Mistral 7B v0.3 Model

5.2.1 Architecture Overview

The mistralai/Mistral-7B-v0.3 is our base model and it represents a significant advancement in transformer architecture design [17]. Traditional Transformer architectures face scalability challenges due to the quadratic computational complexity of their self-attention mechanism with respect to sequence length.[4] Mistral addresses these limitations using grouped query attention, sliding window attention and rotary position embedding.

Grouped-Query Attention (GQA) This technique groups together similar questions the model asks when processing text, instead of handling each question separately. This approach gives almost the same quality results as more expensive methods while using much less computational power [2, 17].

Sliding Window Attention (SWA) Mistral uses a sliding window approach that only looks at a limited number of nearby words at a time instead of looking at every single word in the entire text. This design makes processing long texts much more efficient because the computational work grows linearly instead of exponentially as the text gets longer. [4, 17]

Rotary Position Embedding (RoPE) Mistral uses RoPE to help the model understand word positions in sentences. RoPE tells the model both the exact location of each word and how far words are from each other. This approach allows the model to handle texts of different lengths flexibly, automatically makes words that are far apart have less influence on each other [28].

5.2.2 Technical Specifications

We used "unsloth/mistral-7b-v0.3-bnb-4bit" on our implementation. The model specifications are as follows [17]^{1 2}:

Model	mistralai/Mistral-7B-v0.3	unsloth/mistral-7b-v0.3-bnb-4bit
Source	Official Mistral AI	Unsloth
Parameters	7.25 billion	3.87 billion
Precision	16-bit	4-bit
Quantization	None	BitsAndBytes 4-bit
Architecture	GQA and SWA	GQA and SWA
Memory Usage	High	Reduced (62% less)
Inference Speed	Standard	2.2x faster
Model Quality	Full precision	Minimal degradation

Table 5.2. Comparison of Mistral 7B Model: Original and Optimized

5.3 LLaMA 3.1 8B Model

5.3.1 Architecture Overview

The meta-llama/Llama-3.1-8B model is part of Meta's Llama 3 family and can work with eight different languages while handling very long texts (up to 128,000 tokens) at once [10].

Grouped Query Attention (GQA) Llama 3.1 8B uses Grouped Query Attention with 8 specialized components to make processing more efficient. This design uses less memory when generating text while keeping the same quality performance [10, 2].

Enhanced Tokenization The model uses an expanded vocabulary of 128,000 tokens, which means it can represent text more efficiently than older versions. This allows better text compression without making English processing worse [10].

Rotary Position Embedding (RoPE) Configuration: Llama 3.1 8B adjusts its position understanding system to handle much longer texts by increasing a RoPE hyperparameter [10, 33].

5.3.2 Technical Specifications

Based on our implementation of the unsloth/Meta-Llama-3.1-8B-bnb-4bit which utilizes the meta-llama/Llama-3.1 as the base model, the specifications are as follows^{3 4}:

¹Mistral AI. Mistral-7b-v0.3. <https://huggingface.co/mistralai/Mistral-7B-v0.3>, 2024.

²Unsloth. mistral-7b-v0.3-bnb-4bit. <https://huggingface.co/unsloth/mistral-7b-v0.3-bnb-4bit>, 2024. 4-bit quantized version with BitsAndBytes.

³Meta. Llama-3.1-8b. <https://huggingface.co/meta-llama/Llama-3.1-8B>, 2024.

⁴Unsloth. Meta-llama-3.1-8b-bnb-4bit. <https://huggingface.co/unsloth/Meta-Llama-3.1-8B-bnb-4bit>, 2024. 4-bit quantized version with BitsAndBytes

Model	meta-llama/Llama-3.1-8B	unsloth/Meta-Llama-3.1-8B-bnb-4bit
Source	Official Meta	Unsloth
Parameters	8.3 billion	4.65 billion
Precision	16-bit	4-bit
Quantization	None	BitsAndBytes 4-bit
Architecture	Transformer with GQA	Transformer with GQA
Memory Usage	High	Reduced (58% less)
Inference Speed	Standard	2.4x faster
Model Quality	Full precision	Minimal degradation

Table 5.3. Comparison of Llama 3.1 8B Model: Original and Optimized

5.4 Gemma 2 9B Model

5.4.1 Architecture Overview

The google/gemma-2-9b model is part of Google's Gemma 2 family of lightweight, open language models. The Gemma 2-9B model uses a decoder-only transformer design with several key improvements that make it different from traditional models [30].

Layers and Grouped Query Attention (GQA) The model has 42 layers with specific dimensions, using 16 attention heads with 8 key-value heads in a Grouped Query Attention setup. It uses GQA with 2 groups, which research shows makes the model faster during text generation while keeping the same quality performance [30].

Local Sliding Window and Global Attention Gemma 2 alternates between two types of attention in different layers: local attention that looks at nearby words and global attention that can look at more distant words. This gives the model both detailed local understanding and broader context awareness [30].

Logit Soft-Capping The model uses a technique called logit capping. Logit capping prevents output values from becoming too extreme by applying a mathematical function $logits = softcap * \tanh(logits / softcap)$ on them. This helps keep the values within reasonable bounds and stabilize training and generation [30].

5.4.2 Technical Specifications

Based on our "unsloth/gemma-2-9b-bnb-4bit" implementation, the model specifications are as follows ^{5 6}:

⁵Google. gemma-2-9b. <https://huggingface.co/google/gemma-2-9b>, 2024.

⁶Unsloth. gemma-2-9b-bnb-4bit. <https://huggingface.co/unsloth/gemma-2-9b-bnb-4bit>, 2024. 4-bit quantized version with BitsAndBytes.

Model	google/gemma-2-9b	unsloth/gemma-2-9b-bnb-4bit
Source	Official Google	Unsloth
Parameters	9.24 billion	5.28 billion
Precision	16-bit	4-bit
Quantization	None	BitsAndBytes 4-bit
Architecture	GQA + Interleaved Attention	GQA + Interleaved Attention
Memory Usage	High	Reduced (58% less)
Inference Speed	Standard	2.2x faster
Model Quality	Full precision	Minimal degradation

Table 5.4. Comparison of Gemma 2 9B Model: Original and Optimized

5.5 Yi 1.5 6B Model

5.5.1 Architecture Overview

The unsloth/Yi-1.5-6B-bnb-4bit model is derived from Yi 1.5 6B, a 6-billion parameter bilingual language model created by 01.AI as part of their Yi model family. This Transformer-based architecture was pre-trained on an extensive corpus comprising 3 trillion tokens spanning multiple languages, with specialized capabilities in English and Chinese. The model demonstrates proficiency across diverse natural language processing tasks, encompassing language comprehension, commonsense inference, and reading comprehension applications. The bnb-4bit variant implements 4-bit quantization through the bitsandbytes library, substantially reducing memory overhead from approximately 12GB to under 4GB, thereby facilitating efficient fine-tuning and inference on hardware with constrained computational resources while maintaining competitive performance characteristics [1]^{7 8}

5.6 Qwen2 7B Model

5.6.1 Architecture Overview

The unsloth/Qwen2-7B-bnb-4bit model incorporates several architectural innovations to enhance performance and efficiency. Rather than employing traditional multi-head attention mechanisms, Qwen2 utilizes Grouped Query Attention, which reduces memory overhead associated with key-value caching during inference and substantially improves generation throughput. To enable processing of extended contexts, the architecture implements Dual Chunk Attention, a mechanism that partitions lengthy input sequences into processable segments. This chunking strategy preserves the original attention computation for inputs within chunk boundaries while maintaining effective relative positional encoding both within individual chunks and across chunk boundaries, thereby strengthening long-context modeling capabilities. The model further incorporates YARN to recalibrate attention weights for improved extrapolation to longer sequence lengths. Additional architectural components include SwiGLU activation functions, Rotary Positional Embeddings for position encoding, attention layers with QKV bias terms, and RMSNorm with pre-normalization to ensure training stability. The bnb-4bit quantization applies 4-bit precision reduction via the bitsandbytes library, dramatically decreasing memory requirements to approximately 4GB while preserving model quality for efficient deployment

⁷01.AI. Yi-1.5-6b. <https://huggingface.co/01-ai/Yi-1.5-6B>, 2024

⁸Unsloth. Yi-1.5-6b-bnb-4bit: 4-bit quantized yi-1.5-6b model. <https://huggingface.co/unsloth/Yi-1.5-6B-bnb-4bit>, 2024. 4-bit quantized version using bitsandbytes.

and fine-tuning applications. [26]⁹¹⁰.

5.7 Model Selection Explanation

We utilized the Unsloth Bits and Bytes 4-bit quantized model versions to leverage Unsloth’s optimized fine-tuning capabilities. These models are specifically designed for compatibility with Unsloth’s training framework, enabling efficient LoRA fine-tuning with reduced memory overhead and accelerated training speeds necessary for our model adaptation experiments.

In addition, we used three different models for evaluation of how different architectural approaches, training methodologies, and efficiency optimizations impact political speech generation quality, consistency, and computational requirements.

5.8 Tools and Environment

The following tools and libraries were used for our implementation:

- **Hugging Face Transformers:** Model loading, tokenization, and training loop management.
- **Unsloth:** Lightweight and memory-efficient LoRA-based fine-tuning for LLMs.
- **PyTorch:** Backend for tensor operations and GPU acceleration.
- **Datasets:** Hugging Face’s datasets library was used to manage and preprocess training corpora.
- **AWS:** Training was performed on A100 GPU instances.
- **scikit-learn:** For train-test splitting and basic preprocessing in the classification task.

⁹Qwen Team. Qwen2.5-7b. <https://huggingface.co/Qwen/Qwen2.5-7B>, 2024

¹⁰Unsloth. Qwen2-7b-bnb-4bit: 4-bit quantized qwen2-7b model. <https://huggingface.co/unsloth/Qwen2-7B-bnb-4bit>, 2024. 4-bit quantized version using bitsandbytes

Fine-tuning

6.1 Unsloth Framework

Our model fine-tuning process utilized the Unsloth framework, an optimization toolkit specifically developed for efficient large language model adaptation ¹.

6.2 Parameter-Efficient Fine-tuning with QLoRA

Firstly, we applied Quantized Low-Rank Adaptation (QLoRA) to our model to make it parameter efficient. This implementation integrates 4-bit model quantization with low-rank matrix adaptation for efficient model customization. We chose this approach for our political speech generation because it allows us to adapt large pre-trained models to political speeches without needing massive computational resources or losing the models' general language abilities.

Table 6.1 details the QLoRA configuration parameters applied through Unsloth's `get_peft_model` interface, with justifications for each choice based on Unsloth's finetuning guide ².

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 6.1. Complete QLoRA Configuration Parameters

6.3 Trainer Implementation and Configuration

We fine-tuned our models using Supervised Fine-Tuning on instruction-response pairs, formatted using model-specific chat templates. We chose this approach because our models needed to learn how to respond with political speeches based on provided questions and information (political affiliation, EuroVoc topic, house, debate section, political orientation). More specifically, we used SFTTrainer, which is designed for training models on conversation-style data to learn how to follow instructions,

¹D. Han, M. Han, and Unsloth team. Unsloth: Fast, memory-efficient fine-tuning of LLMs. GitHub repository. <https://github.com/unslothai/unsloth>, 2023

²Unsloth. Fine-tuning pretrained language models (unsloth llms guide). <https://docs.unsloth.ai/get-started/fine-tuning-llms-guide>, 2025.

and QLoRA, which updates only a small part of the model instead of changing it entirely in order to be more computationally efficient.

6.3.1 Multi-Model Chat Template Architecture

System Prompt All models are trained with the following template:

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

Each model architecture wraps this shared structure in its own special formatting tokens and control sequences.

Mistral Model Template:

```
<s>[INST] SYSTEM_PROMPT
Context: {context}
Instruction: {instruction} [/INST]
```

Llama 3.1 Model Template:

```
<|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|>
```

Gemma 2 Model Template:

```
<bos><start_of_turn>user
SYSTEM_PROMPT
Context: {context}
Instruction: {instruction}<end_of_turn>
<start_of_turn>model
```

Gwen2 Model Template (ChatML format):

```
<|im_start|>system
SYSTEM_PROMPT<|im_end|>
<|im_start|>user
Context: {context}
Instruction: {instruction}<|im_end|>
<|im_start|>assistant
```

Yi-1.5 Model Template (ChatML format):

```
<|im_start|>system
SYSTEM_PROMPT<|im_end|>
<|im_start|>user
Context: {context}
Instruction: {instruction}<|im_end|>
<|im_start|>assistant
```

The instruction field includes the question prompt or the generic instruction when there are no prompts available. The context fields includes the following organized political information:

- **PARTY:** Political party affiliation (eg. Conservative)
- **EUROVOC TOPIC:** EuroVoc classification (eg. TRADE)
- **SECTION:** Parliamentary debate section
- **POLITICAL ORIENTATION:** Political orientation label (eg. Right)
- **HOUSE:** House of Commons or House of Lords

This setup helps fine-tuned models learn to give responses that match a party's views, stay on topic, follow parliamentary rules and political views.

6.3.2 Training Arguments and Hyperparameter Selection

Table 6.2 details the parameters we used, and as recommended by the Unsloth finetuning guide, we kept the default settings.

Parameter	Value	Justification
Batch Size	64	GPU memory optimization
Learning Rate	2e-4	Standard for LoRA fine-tuning
Max Steps	11194	2 epochs
Warmup Steps	336	10% of max steps for stability
Optimizer	adamw	Memory-efficient
Weight Decay	0.01	Prevents overfitting on political data
Max Sequence Length	1024	Optimal value for our train dataset median length
Scheduler	linear	Linear learning rate schedule

Table 6.2. Complete Training Configuration for Parliamentary Speech Generation

The training configuration maintains consistency across all model architectures and political contexts, with automated data loading, standardized train-test splits (80-20), and comprehensive resource management including GPU cache clearing between sessions and monitoring of computational efficiency metrics across Google Colab A100 instances.

6.3.3 Training Pipeline

Training Strategy We train each model separately, creating individual result folders named `unified_model_name`. We use the same data preparation steps for every dataset to ensure consistency. Each dataset is split into 80% for training and 20% for testing using the same random seed, so results can be reproduced. The pipeline handles different data formats automatically, extracting political information like party affiliation, topics, and parliamentary sections, and provides backup instructions when the original instruction is missing. Each training session uses the appropriate model's chat format.

Files Inside Each Model Folder After the training is complete, we save the LoRA adapter weights (PEFT), not the full model. The weights are stored in `./trained_models/MODEL_NAME/`

Based on our code and LoRA training, each folder contains:

- `adapter_model.bin` - The actual LoRA adapter weights
- `adapter_config.json` - Configuration for the LoRA adapter (rank, alpha, target modules, etc.)
- `tokenizer.json` - Tokenizer files
- `tokenizer_config.json` - Tokenizer configuration
- `special_tokens_map.json` - Special token mappings

Speech Generation

7.1 Generation Pipeline Architecture

We built a speech generation system that loads our trained models and creates political speeches based on a generated input. The system accepts multiple types of input including instruction prompt, EuroVoc topic, section, political party affiliation, house and political orientation.

7.2 Input Mapping

To ensure consistency and fairness across model comparisons, we generated a unified set of input sections and prompts from the test set, which guaranteed that all models received identical generation tasks. We maintained realistic distributional characteristics by sampling houses according to their actual representation in the UK Parliament (78% House of Commons, 22% House of Lords) and political parties according to their respective weights in the dataset (Conservative: 59%, Labour: 24%, Scottish National Party: 5%, and so forth). Each party was associated with its corresponding political orientation (e.g., Conservative as Centre-right, Labour as Centre-left), and we enforced valid house-party mappings by restricting certain parties to their appropriate chambers (e.g., Bishops only in House of Lords). The instruction format comprised 10% task-specific prompts randomly selected from the unseen prompt set and 90% generic instructions following a standardized template. We additionally mapped the sections and prompts to EuroVoc topics using Kevlar to ensure thematic consistency and prevent misaligned inputs. This controlled generation approach, with all inputs pre-saved in a shared JSON file, ensured that any observed performance differences could be attributed solely to model capabilities rather than variations in input distribution or prompt complexity.

Party	Weight	Political Orientation
Conservative	0.59	Centre-right
Labour	0.24	Centre-left
Scottish National Party	0.05	Centre-left
Liberal Democrats	0.05	Centre to centre-left
Crossbench	0.028	Unknown
Democratic Unionist Party	0.016	Right
Independent	0.010	Unknown
Plaid Cymru	0.006	Centre-left to left
Green Party	0.005	Left
Non-Affiliated	0.003	Unknown
Bishops	0.002	Unknown

Table 7.1. Party Distribution and Political Orientation

House	Proportion
House of Commons	0.78
House of Lords	0.22

Table 7.2. *House Distribution in the Dataset*

7.3 Model Loading and Inference Configuration

For the fine-tuned models, the system loads the models and applies the trained LoRA adapters using the PEFT library. The pipeline detects LoRA adapter files by checking for `adapter_config.json` and loads the corresponding model before applying the political speech fine-tuning weights. The system automatically determines the appropriate device mapping strategy: single-GPU mode forces allocation to device 0 when `CUDA_VISIBLE_DEVICES` is set, while multi-GPU environments use automatic device placement with `device_map="auto"`. All models are loaded with `torch.float16` precision on GPU or `torch.float32` on CPU, with `trust_remote_code=True` to support custom model implementations. For baseline models, the system downloads pre-trained 4-bit quantized versions from Unsloth’s HuggingFace repository (`unsloth/mistral-7b-v0.3-bnb-4bit`, etc.) without applying any adapters, enabling direct comparison between pre-trained and fine-tuned variants.

7.4 Generation Parameters and Decoding Strategy

Following best practices for neural text generation, we implemented nucleus sampling (top-p) as our primary decoding strategy. We chose this over traditional decoding methods like greedy search and beam search as they often produce repetitive or incoherent text despite achieving low perplexity [13].

We configured generation parameters specifically for political speech coherence and diversity.

Parameter	Value / Description
SPEECHES_PER_MODEL	2700 speeches generated per model.
TEMPERATURE	0.7 — balances coherence and lexical variation in political speech generation.
TOP_P	0.85 — applies nucleus sampling for focused yet diverse outputs.
REPETITION_PENALTY	1.2 — penalizes redundant phrasing for stylistic diversity.
BATCH_SIZE	32 — generates in batches for approximately 3 times speed improvement.
ENABLE_BATCHING	True — enables batched generation; when False, generation is sequential.
MIN_WORD_COUNT	43 — lower quality threshold (P10) to filter abnormally short speeches.
MAX_WORD_COUNT	635 — upper quality threshold (P90) to filter overly long speeches.
MAX_NEW_TOKENS	850 — cap on generated length (= 1.33 times P90 speech length).
GENERATION_SEED	42 — ensures reproducibility across runs.

Table 7.3. *Model Generation Configuration Parameters*

7.5 Multimodal Chat Template

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

System Prompt

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 parliamentary language appropriate for the specified House.
The speech should reflect the political orientation and typical positions of the specified party on the given topic.

Model Templates

Each model architecture wraps this shared structure in its own special formatting tokens and control sequences.

Mistral Model Template:

```
<s>[INST] SYSTEM_PROMPT
Context: {context}
Instruction: {instruction} [/INST]
```

Llama 3.1 Model Template:

```
<|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|>
```

Gemma 2 Model Template:

```
<bos><start_of_turn>user
SYSTEM_PROMPT
Context: {context}
Instruction: {instruction}<end_of_turn>
<start_of_turn>model
```

Gwen2 Model Template (ChatML format):

```
<|im_start|>system
SYSTEM_PROMPT<|im_end|>
<|im_start|>user
Context: {context}
Instruction: {instruction}<|im_end|>
<|im_start|>assistant
```

Yi-1.5 Model Template (ChatML format):

```
<|im_start|>system
SYSTEM_PROMPT<|im_end|>
<|im_start|>user
Context: {context}
Instruction: {instruction}<|im_end|>
<|im_start|>assistant
```

7.6 Speech Validation

We implemented a `SpeechValidator` class to ensure the quality, coherence, and relevance of generated parliamentary speeches. It performs systematic checks to detect common issues arising in generated outputs, including template artifacts, encoding errors, language mismatches, degenerative repetition, off-topic content, and structural inconsistencies. Generated speeches underwent a 9-step validation procedure:

1. **Template Marker Detection:** Detects 27 template artifacts including role markers (`\nuser`, `\nassistant`), context labels (`Context:`, `Instruction:`), and special tokens (`<|im_start|>`, `[INST]`, etc.).
2. **Unicode Corruption Detection:** Identifies 14 specific corruption patterns (`U+03B2`, `U+258D`, `<2mass>`) and checks 11 forbidden Unicode ranges including CJK (Chinese, Japanese, Korean), Cyrillic, Arabic, Thai, and technical symbol blocks.
3. **Language Detection:** Uses `spacy-langdetect` with the `en_core_web_sm` model to detect non-English text. Only performed on texts ≥ 30 characters, flagging languages detected with $> 85\%$ confidence.
4. **Repetition Detection:** Detects three repetition patterns: (1) same word repeated > 3 times consecutively, (2) sequences of 3-7 words repeated > 3 times, (3) counting patterns with > 5 ordinal words (first, second, third, etc.).
5. **Semantic Relevance Check:** Computes cosine similarity between speech embeddings and context description ("UK parliamentary debate about {section} on {topic}") using ('all-MiniLM-L6-v2'). Flags speeches with similarity < 0.08 as off-topic.
6. **Length constraints:** Validates word count (43-635).
7. **Concatenation detection:** Detects concatenated speeches (≥ 4 opening phrases like "My Lords" or "Mr Speaker")
8. **Corrupted endings detection:** Identifies nonsensical endings (`U+258D`).
9. **Refusal pattern matching detection:** Catches AI refusal patterns ("I cannot generate", "I'm sorry but I cannot").

The `SpeechValidator` class ensures that outputs:

- Are free of prompt/template/meta-commentary/HTML tag prefixes, suffixes and artifacts
- Contain valid English text (using `spacy-langdetect` with 0.85 confidence threshold)

- Avoid degenerate repetitions (max 3 consecutive word repeats, sequences of 3-7 words)
- Remain contextually relevant (semantic similarity ≥ 0.08 using all-MiniLM-L6-v2)
- Follow expected structural conventions (43-635 words, proper punctuation)

7.7 Examples of Generated Speeches

Scottish National Party - Gemma Model

ENERGY Domestic Renewable Energy

House of Commons Centre-left

"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!

Table 7.4. *Baseline Model Speech - Gemma*

Conservative Party - Llama Model

INDUSTRY Hawk Aircraft: Overseas Promotion

House of Commons Centre-right

"Address the debate on Hawk Aircraft: Overseas Promotion on INDUSTRY."

Generated Response: Text: "The government's recent decision to promote exports of our world-renowned aircraft manufacturing capabilities is both commendable and timely. As an advocate for economic growth and international trade, I fully support this initiative. "Overseas promotion plays a crucial role in boosting Britain's economy by increasing export revenue. It also helps create jobs domestically while enhancing our nation's reputation as a global leader in innovation and engineering excellence. "Hawk Aircraft has long been synonymous with quality craftsmanship, safety standards, and technological advancements that set it apart from its competitors worldwide. By highlighting these strengths during overseas promotional campaigns, we can secure lucrative contracts abroad, ensuring sustained employment opportunities here at home. "The current market trends indicate growing demand for advanced fighter jets across various regions globally. With our exceptional track record, skilled workforce, and robust infrastructure supporting research & development initiatives - now more than ever before - there exists ample scope for us to capitalise upon emerging prospects beyond Europe too! "In conclusion then; let me urge my esteemed colleagues within parliament today to rally behind efforts aimed towards fostering increased foreign sales of Hawks through strategic outreach programmes designed specifically tailored according to individual markets' needs thereby contributing significantly toward bolstering overall industrial output alongside creating much needed job openings amidst ongoing challenges faced due COVID19 pandemic."

Table 7.5. *Baseline Model Speech - Llama*

Conservative Party - Mistral Model

ENVIRONMENT The Situation in the Gulf

House of Commons Centre-right

"Address the debate on The Situation in the Gulf on ENVIRONMENT."

Generated Response: The Conservatives have long supported the EU's position as an independent global player, and we believe that this is particularly important in the context of environmental policy. We recognize the importance of working with our European partners to address issues such as climate change and biodiversity loss, but also acknowledge the need to work closely with countries outside Europe to achieve these goals. This is especially true when it comes to addressing the challenges posed by pollution, deforestation, and other threats to our planet. As such, I would like to take this opportunity to reiterate my support for the EU's commitment to leading the way on environmental protection and call upon all Member States to join us in making sure that the Gulf region remains safe from further damage caused by human activities. We must ensure that any trade agreements between the EU and its trading partners include strong provisions for protecting the environment, including measures to reduce greenhouse gas emissions and preserve natural resources. Additionally, we should continue to encourage dialogue with non-EU states on matters related to sustainable development so that they can benefit from our experience while avoiding potential conflicts over resource allocation or access rights. Finally, I urge my colleagues here today to consider how best we can leverage existing partnerships within both the GCC and Asia Pacific regions towards achieving greater cooperation across borders when tackling shared problems affecting regional ecosystems and economies alike.

Table 7.6. *Baseline Model Speech - Mistral*

Conservative Party - Qwen Model

POLITICS Right-wing Extremism

House of Commons Centre-right

"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.

Table 7.7. *Baseline Model Speech - Qwen*

Labour Party - Yi Model

AGRI-FOODSTUFFS Food Waste

House of Commons Centre-left

"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.

Table 7.8. *Baseline Model Speech - Yi*

Conservative Party - Gemma Model**LAW HM Courts and Tribunal Service****House of Commons Centre-right****"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. That includes dealing with sexual offences cases through sex offender treatment programmes, domestic violence offenders going back time after time before magistrates' courts because their partners just cannot get away from them, complex fraudsters whose activities go beyond anything else happening locally within any other single jurisdiction area including county lines drug dealers working out whether someone needs help rather than prosecution if addiction problems make sure she goes down another route towards recovery instead. We believe strongly too much focus remains placed solely onto sentencing outcomes without enough attention paid equally important issues relating thereto during trial proceedings themselves. First though let me remind you why reform matters here. Firstly unlike most European countries ours still uses outdated technology often reliant largely manual processes. Secondly despite having fewer judges per capita compared internationally costs continue rising annually partly due inefficient use resources. Thirdly although overall number defendants decreasing significantly certain categories remain stubbornly high eg serious violent crime. Fourthly finally while public safety obviously paramount nobody wants see unnecessary burden place taxpayers pockets unnecessarily either therefore government determined find balance between protecting society ensuring value money spent means delivering better services faster cheaper way. Let us take look some specific examples first. One key element behind new model involves increasing efficiency reducing bureaucracy associated running separate local offices spread throughout entire country. Instead proposed structure based regional centres supported centralised national hub coordinating activity nationwide. This approach allows greater economies scale achieved previously combined together creating potential savings estimated somewhere region £1billion pound sterling equivalent approximately four thousand million euros. Using latest digital platforms makes communication quicker easier making whole process smoother less error prone resulting improved productivity whilst simultaneously minimising delays caused traditional paper systems. Finally introducing common standards best practice enables consistency applied uniformly regardless location thereby eliminating inconsistencies arising disparate approaches taken independently hitherto. So far successful implementation seen already evidenced positive results observed thus far namely reduced waiting times decreased costings increased accuracy along enhanced customer satisfaction rates. Overall outcome represents major step forward improving accessibility quality service delivery provided universally everywhere. Furthermore alongside rationalisation programme announced earlier week aimed optimising capacity utilisation maximising output via consolidation existing stock portfolio offers further scope improvement performance even higher level. Given foregoing factors outlined above clear message sent loud strong terms indeed comes sense urgency required address shortcomings highlighted herein order meet demands ever growing complexity modern life successfully hence urgent action needed now asap please? In conclusion reforms represent significant investment designed bring benefits both users providers alike. Therefore vital importance securing necessary buy-in stakeholder groups concerned affected parties ensures smooth transition period ahead. Otherwise risk falling short expectations set beforehand ultimately leading potentially catastrophic consequences detrimental effects long term sustainability itself. Question Resolved, That this House recognises Her Majesty's Government's duty under section 2(7) of the Justice.

Table 7.9. *Finetuned Model Speech - Gemma*

Conservative Party - Llama Model**INTERNATIONAL ORGANISATIONS UK-EU Trade Policy Options****House of Commons Centre-right****"Address the debate on UK-EU Trade Policy Options on INTERNATIONAL ORGANISATIONS."**

Generated Response: I am grateful to my right hon. Friend for his response. As he will be aware, I have long campaigned against our membership of the European Union because we were told that it was not possible to leave without damaging ourselves economically; yet here we are with an economy growing faster than any other major advanced country's. Does this not prove that those who warned us about economic damage if we left were wrong? Will my right hon. Friend assure me that when negotiations start there will be no question but that we must take back control over laws relating to trade policy?

Table 7.10. *Finetuned Model Speech - Llama***Conservative Party - Mistral Model****INTERNATIONAL RELATIONS Nuclear Weapons (International Relations Committee Report)****House of Commons Centre-right****"Address the debate on Nuclear Weapons (International Relations Committee Report) on INTERNATIONAL RELATIONS."**

Generated Response: 2018 is a year of major anniversaries in our history of nuclear weapons. It marks 70 years since atomic bombs were dropped on Japan; it also marks 70 years since NATO was established—I am sure that will be remembered by many Members on both sides of this House; and it also marks 70 years since the UN General Assembly adopted its first resolution calling for an end to nuclear testing. In addition, we mark this year the start of negotiations between President Reagan and Mikhail Gorbachev towards reducing nuclear stockpiles I commend my noble Friend Lord Howell for his work as Chair of the International Relations Select Committee, which produced the report on nuclear deterrence, "Time to Stop the Clock", earlier this year. He said at the time that he hoped that the Government would give serious consideration to all aspects of this important issue. As we have heard today from him and others who spoke eloquently about these issues during this debate, the committee's report contains several key recommendations that I hope the Minister will respond to. These include calls to support multilateral disarmament efforts, particularly through the conference on the humanitarian impact of nuclear weapons in December; to ensure that any future strategic defence review includes proposals to reduce the number of operational warheads carried by each Trident submarine, not just those stored separately ashore; to increase funding for the alternative delivery vehicle programme; to continue to engage with other states such as China, India and Pakistan to encourage them to join the global norm against the use or threat of use of nuclear weapons; and to provide regular updates to Parliament on progress made on implementing the Government's non-proliferation treaty obligations.

Table 7.11. *Finetuned Model Speech - Mistral***Labour Party - Gwen Model****POLITICS Hong Kong Arrests Under National Security Law****House of Lords Centre-left****"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. We know from Amnesty International and Human Rights Watch reports that the arrests show the full extent of the chilling effect of the draconian new laws passed by Beijing in July, which threaten freedom of expression, association and peaceful protest among other things. These latest developments come after a year marked by escalating threats to civil liberties, including widespread police brutality during pro-democracy protests across Hong Kong. As well as the detention of Ms Wang, there was also the arrest earlier this month of Joshua Wong—another young activist who had recently turned down a scholarship at Harvard University because he believed that education could not be separated from politics. Are the Government concerned that these latest arrests demonstrate continued disregard for basic freedoms?

Table 7.12. *Finetuned Model Speech - Gwen*

Scottish National Party - Yi Model INTERNATIONAL RELATIONS Flooding: EU Solidarity Fund House of Commons Centre-left "Address the debate on Flooding: EU Solidarity Fund on INTERNATIONAL RELATIONS." <i>Generated Response:</i> 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.
--

Table 7.13. *Finetuned Model Speech - Yi*

Evaluation Framework

8.1 Evaluation Methodology Overview

We implement an evaluation framework to assess how well our trained political speech models work. Our evaluation uses the following methods: language quality (using perplexity, Distinct-N, and Self-BLEU), whether the content makes sense (semantic similarity using GRUEN, BERTScore, and MoverScore), whether the writing style sounds like the political parties and orientation (style and ideology analysis using word embeddings), and an overall evaluation using an LLM-as-a-Judge.

8.2 Linguistic Quality and Diversity Metrics

8.2.1 Perplexity

Perplexity measures how surprised a language model is when it encounters specific text. Lower perplexity scores indicate text that seems natural and predictable to the model, while higher scores suggest text that appears unusual or artificial. In simple terms, perplexity indicates whether AI-generated political speeches sound human-like or robotic [16].

Perplexity Implementation

We used the official HuggingFace library to calculate perplexity for our political speeches. HuggingFace provides a standard, well-tested implementation that is widely used in research. This ensures our results are reliable and comparable to other studies. We use the `evaluate.load("perplexity", module_type="metric")` function, which provides access to the HuggingFace's perplexity. Input texts are processed in batches and are tokenized using the `AutoTokenizer`. We also make use of manual cross-entropy computation with shifted logits and labels. Our analysis uses GPT-2 base with 117 million parameters as the evaluation model. GPT-2 was trained on a large variety of text from the internet, giving it broad knowledge of how natural English text should look. For our evaluation, it processes a maximum of 512 words per speech, 8 speeches simultaneously in batches, and includes automatic start token addition.

8.2.2 Distinct-N

Distinct-N metrics evaluate lexical diversity by measuring the ratio of unique n-grams to the total number of tokens in the generated text. Higher Distinct-N scores indicate more diverse vocabulary and less repetitive patterns, while lower scores suggest frequent repetition and more formulaic language generation. We compute scores individually for each speech [22].

Distinct-N Implementation

We implement Distinct-1, Distinct-2, Distinct-3, and Distinct-4 using the standard sentence-level formulation to assess vocabulary richness and repetition patterns in political speeches. Following the established NLTK implementation, we compute Distinct-N as the ratio of unique n-grams to the total number of tokens in the text [22]. We process speech texts through several preprocessing stages: lowercasing, whitespace normalization, removal of punctuation (retaining only word characters, spaces, and apostrophes), and tokenization with filtering of short tokens ($\text{length} \leq 1$) and pure numeric strings. For each speech, we calculate four distinct metrics using the standard formulation: $\text{Distinct-N} = \frac{|\text{unique n-grams}|}{\text{total tokens}}$ [22]. Specifically, Distinct-1 measures unique unigrams normalized by text length, capturing basic lexical diversity. Distinct-2 measures unique bigrams normalized by text length, reflecting phrase-level variety. Distinct-3 and Distinct-4 similarly measure unique trigrams and 4-grams normalized by text length, revealing multi-word phrase patterns and sentence-level structural choices that indicate sophisticated language use.

8.2.3 Self-BLEU

Self-BLEU measures the similarity between generated texts from the same model. Unlike regular BLEU which compares generated text to reference text, Self-BLEU compares each generated text to all other generated texts in the same collection (model). This tells us whether a model is producing varied content or simply repeating similar phrases and ideas. Lower Self-BLEU scores indicate higher diversity (good), while higher scores suggest repetitive content (potentially problematic) [39]¹.

Self-BLEU Implementation

We implemented Self-BLEU using the methodology from DigitalOcean’s automated metrics tutorial. The approach works by taking each generated speech and calculating its BLEU score against every other speech in the same model’s generated collection with the same inputs (party, house, orientation, topic), then averaging all these scores [39].

8.3 Semantic Coherence and Text Quality Evaluation

8.3.1 GRUEN Score

We implement the GRUEN (Grammaticality, non-Redundancy, focUs, structurE, and coNherence) evaluation framework to assess generated political speech quality. We utilize three core components: Grammaticality, Non-Redundancy (as penalty), and Focus (as penalty), which combine into a unified quality metric specifically designed for evaluating neural text generation [38].

Text Preprocessing Following the original GRUEN methodology, we preprocess texts through multiple normalization steps: converting paragraph breaks and newlines to sentence boundaries, normalizing multiple periods, handling decimal numbers (replacing matches like "5. 3" with "UNK" to avoid sentence splitting), and standardizing whitespace. We then tokenize into sentences using NLTK’s sentence tokenizer and filter out sentences with fewer than two non-punctuation words to ensure meaningful grammatical evaluation [38].

¹DigitalOcean Community. Automated metrics for evaluating generated text. <https://www.digitalocean.com/community/tutorials/automated-metrics-for-evaluating-generated-text>

Grammaticality Component The grammaticality component combines two complementary neural approaches:

BERT Masked Language Model: We employ bert-base-cased to compute perplexity scores at the sentence level, measuring how "surprising" the text is to the model (lower perplexity indicates more grammatical text). We process each sentence independently and average the perplexity scores across the entire text.

CoLA Classifier: We use the andeskiyl/bert-base-cased-cola model to predict grammatical acceptability, providing probability scores between 0 and 1 (higher values indicate greater grammatical acceptability).

The final grammaticality score combines these components using the original paper's formulation:

$$G = a_1 \cdot \exp(-\beta_1 \cdot PPL) + a_2 \cdot CoLA \quad (8.1)$$

where $a_1 = 1$, $\beta_1 = 0.5$, $a_2 = 1$, PPL is the BERT perplexity score, and $CoLA$ is the grammatical acceptability score. The exponential transformation $\exp(-\beta_1 \cdot PPL)$ converts high perplexity (indicating ungrammatical text) into low scores, while equal weights ensure both components contribute equivalently. We then rescale it to $[0, 1]$ by scaling it down by factor of 8, shifting the distribution upward by 0.5, and flooring to 0 to prevent negative values [38].

Non-Redundancy Component (Penalty) Non-redundancy measures lexical and phrasal diversity using string-based similarity metrics between consecutive sentences. Following the original implementation, we compute three complementary metrics:

Longest Common Subsequence (LCS): Measures the longest sequence of words appearing in order in both sentences, normalized by sentence length.

Edit Distance: Computes the Levenshtein distance (minimum character-level edits) between sentences, normalized by the maximum sentence length.

Word Overlap: Calculates the proportion of shared words between consecutive sentences using set intersection.

These metrics are averaged across all consecutive sentence pairs to produce a redundancy score, which is then converted to a penalty term. Higher redundancy (lower diversity) results in larger penalties that reduce the final GRUEN score [38].

Focus Component (Penalty) Focus measures topical coherence by assessing semantic consistency across sentences. We employ two complementary semantic similarity approaches:

Word Mover's Distance (WMD): Computes the semantic distance between consecutive sentences using word embeddings from spaCy's `en_core_web_md` model. WMD measures the minimum "travel distance" required to transform one sentence into another in the embedding space.

SpaCy Semantic Similarity: Calculates cosine similarity between sentence embeddings generated by spaCy's neural network.

These metrics are averaged across all consecutive sentence pairs. Lower semantic coherence (indicating topic drift) produces larger penalties that reduce the final GRUEN score [38].

Final GRUEN Score The final score combines grammaticality and penalties using the original formulation:

$$GRUEN = \min\left(1, \max\left(0, G + R_{\text{penalty}} + F_{\text{penalty}}\right)\right) \quad (8.2)$$

where G is the grammaticality score (positive contribution), while R_{penalty} and F_{penalty} are negative terms that reduce the score when redundancy or focus issues are detected. The min-max clamping ensures scores remain in the $[0, 1]$ range [38].

8.3.2 BERTScore

BERTScore uses a pre-trained BERT model to measure how similar two texts are in meaning. Unlike older methods that just count matching words, BERTScore understands what the text actually means and the context around words. In our research, BERTScore is used to evaluate how realistic speeches language models generate, by measuring the semantic similarity between $N=6$ real ParlaMint-GB dataset speeches and the speech the model generated. The evaluation uses the same methodology as research and HuggingFace, providing precision, recall, and F1 scores that indicate how well each model understands and generates political speeches [34].

BERTScore Implementation

BERTScore was computed using the official HuggingFace evaluate library implementation [34]². The metric was loaded using `evaluate.load("bertscore")` with the following configuration:

- `lang="en"`: Specified English language for automatic model selection
- `model_type=None`: Enabled automatic selection of the recommended model for English (RoBERTa-large [34])

8.3.3 MoverScore

MoverScore Implementation

We implement MoverScore following the standard methodology to measure semantic similarity between generated political speeches and real reference speeches. MoverScore quantifies the semantic distance by computing the optimal transport cost required to transform one text's embedding distribution into another's, leveraging contextualized word embeddings and inverse document frequency (IDF) weighting [36]. For each generated speech, we selected the $N=6$ most semantically similar speeches from the original dataset as references, measured by cosine similarity between speech embeddings and the corresponding generation context prompt embedding.

Model Architecture and Embedding Generation We employ `distilbert-base-uncased` to generate contextual embeddings for both reference and generated speeches. The model processes tokenized text through its transformer architecture, and we extract embeddings from the final hidden layer. Text preprocessing includes truncating speeches to 2000 characters and limiting tokenized sequences to 512 tokens (including [CLS] and [SEP] special tokens) to comply with model constraints.

IDF Computation and Weighting We compute corpus-level IDF weights using the standard formulation: $\text{IDF}(w) = \log\left(\frac{N+1}{df(w)+1}\right)$, where N is the total number of documents and $df(w)$ is the document frequency of word w . IDF weights are calculated separately for reference and generated speech corpora using multiprocessing for efficiency, then applied to token embeddings to emphasize semantically important words while down-weighting common terms.

²Hugging Face. Evaluate: A library for easily evaluating machine learning models and datasets. <https://github.com/huggingface/evaluate>, 2022

Token Filtering Before computing similarity, we filter three categories of tokens to focus on meaningful semantic content:

- Punctuation marks (using Python’s `string.punctuation`)
- BERT subword tokens (marked with ‘##’ prefix)
- Stop words (optional parameter, empty by default in our implementation)

Filtered tokens have their embeddings zeroed and IDF weights set to zero, effectively excluding them from the optimal transport calculation.

Earth Mover’s Distance Calculation We compute the Earth Mover’s Distance (EMD) using the Python Optimal Transport (POT) library’s `ot.emd()` function. For each reference-generated pair:

1. Concatenate reference and generated embeddings into a unified embedding matrix
2. Normalize embeddings to unit length: $\hat{e} = \frac{e}{\|e\| + \epsilon}$ where $\epsilon = 10^{-30}$
3. Compute pairwise L2 distances between all token embeddings using batched operations
4. Construct probability distributions c_1 and c_2 from IDF-weighted tokens, normalized to sum to 1
5. Solve the optimal transport problem: $\text{flow} = \text{EMD}(c_1, c_2, D)$ where D is the distance matrix
6. Compute final score: $\text{MoverScore} = \frac{1}{1 + \sum_{i,j} \text{flow}_{i,j} \times D_{i,j}}$

The scoring formula transforms the total transport cost into a similarity measure between 0 and 1, where higher scores indicate greater semantic alignment between reference and generated text.

8.4 Political Spectrum and Party Alignment

We developed an analysis system to examine how models write political speeches, focusing on both political party style and political orientation content [27, 31]. The system reads speech texts and measures various characteristics to understand the linguistic patterns and political positioning of the model’s generated content. This method is based on semantic similarity measurements between generated content and real parliamentary speeches from the training corpus. Our methodology employs two complementary metrics: Political Spectrum Alignment (PSA) and Party Alignment, both computed using dense vector representations derived from a pre-trained sentence transformer model (all-mpnet-base-v2). This approach enables us to evaluate whether the fine-tuned models successfully learned to generate politically coherent content that aligns with the specified political orientations and party affiliations provided in the input prompts.

The evaluation process begins by constructing reference embeddings from authentic parliamentary speeches in the training dataset. For Political Spectrum Alignment, we group all real speeches according to their political orientation codes (ranging from Far-left to Far-right on a 13-point scale) and compute centroid embeddings for each orientation category by averaging the sentence embeddings of all speeches within that category. Similarly, for Party Alignment evaluation, we create party-specific centroid embeddings by grouping real speeches by party affiliation and computing their average representations. These reference centroids serve as archetypal representations of genuine political discourse for each orientation and party, capturing the distinctive linguistic patterns, rhetorical strategies, and thematic emphases characteristic of different political positions.

To evaluate each generated speech, we encode it using the same sentence transformer model and compute its cosine similarity to all reference centroids. For PSA scoring, we identify the closest political orientation centroid and calculate an alignment score based on the ideological distance between the expected orientation (specified in the generation prompt) and the closest matching orientation. The PSA score ranges from 0 to 1, where higher scores indicate that the generated speech semantically aligns with content from the intended political position. Specifically, the score is computed as $PSA = \text{cosine_similarity}(\text{speech}, \text{embedding}) * \max(0, 100 - \frac{d}{12} \times 100)$, where d represents the absolute distance between the expected and predicted orientations on the political spectrum scale. For Party Alignment, we compute the cosine similarity between the generated speech embedding and the expected party's centroid embedding, on a 0-1 scale. This metric assesses whether the model successfully captures party-specific linguistic characteristics beyond general ideological positioning.

8.5 LLM-as-a-Judge Evaluation

We implemented an 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.

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.

8.5.1 Metric 1: Coherence

Evaluation Criteria

Does the speech follow a clear logical progression? Are arguments well-connected and ideas flow naturally with appropriate parliamentary structure?

Prompt Template and Scoring Rubric

GOAL

You are a senior parliamentary discourse analyst with 15 years of experience evaluating political speech quality in Westminster-style parliaments. Your task is to evaluate the coherence (logical flow and structure) of this UK parliamentary speech.

OUTPUT

Below is the parliamentary speech to evaluate:

```
<output>
{{ANSWER}}
</output>
```

EVALUATION CRITERIA AND SCORING RUBRIC

```
<evaluation_criteria>
```

Does the speech follow a clear logical progression? Are arguments well-connected? Is parliamentary structure maintained?

```
</evaluation_criteria>
```

```
<scoring_rubric>
```

- Score 1-2: Incoherent. No discernible logical structure, fragmented or nonsensical.
- Score 3-4: Poor coherence. Disjointed arguments, unclear progression, difficult to follow main point.
- Score 5-6: Adequate coherence. Main argument identifiable but weak connections, some repetition or tangents.
- Score 7-8: Strong coherence. Mostly logical progression with minor transition issues, follows parliamentary structure.
- Score 9-10: Exceptional logical flow. Each point builds naturally, clear thesis and conclusion, perfect parliamentary structure.

```
</scoring_rubric>
```

INSTRUCTIONS FOR THE EVALUATION

1. Before scoring, analyze the speech structure step-by-step.
2. Quote specific evidence or patterns from the speech supporting your assessment.
3. Use the full 1-10 range - don't default to middle scores (5-7) without justification.
4. Pay attention to how well arguments connect and whether parliamentary conventions are maintained.

FORMAT FOR THE EVALUATION

- Write your verbal feedback inside <feedback> tags without any surrounding text.
- Write the numeric score inside <score> tags, always after the feedback.

Please accurately evaluate the speech. Strictly adhere to the evaluation criteria and rubric.

8.5.2 Metric 2: Conciseness

Evaluation Criteria

Is the message conveyed efficiently without excessive verbosity, within the context of formal parliamentary speech where longer discourse is expected?

Prompt Template and Scoring Rubric

GOAL

You are a senior parliamentary discourse analyst with 15 years of experience evaluating political speech quality. Your task is to evaluate the conciseness of this UK parliamentary speech using parliamentary standards (where formal speeches are naturally longer than everyday communication).

OUTPUT

Below is the parliamentary speech to evaluate:

<output>

{{ANSWER}}

</output>

EVALUATION CRITERIA AND SCORING RUBRIC

<evaluation_criteria>

Does the speech convey its message efficiently without unnecessary verbosity? Remember: Parliamentary speeches are naturally more formal and verbose - assess within this context.

</evaluation_criteria>

<scoring_rubric>

- Score 1-2: Extremely verbose. Main argument lost among unrelated statements, excessive repetition, rambling.
- Score 3-4: Verbose. Significant digression into unrelated matters, could be 30-50% shorter while preserving intent.
- Score 5-6: Adequate. Main point clear but diluted by length or tangential references, some ceremonial language adds weight.
- Score 7-8: Concise. Clear focus with minor digressions (constituency examples), formal niceties don't impede core message.
- Score 9-10: Highly concise for parliamentary standard. Focused and disciplined, each point supports main argument, courtesy phrases used sparingly.

</scoring_rubric>

INSTRUCTIONS FOR THE EVALUATION

1. Identify the main argument and check for focus vs digressions.
2. Assess if courtesy and formality serve rhetorical purpose (parliamentary context).

3. Detect unnecessary repetition or verbose phrasing.
4. Remember: Parliamentary speeches are naturally longer than everyday communication - adjust expectations accordingly.
5. Quote examples showing appropriate length or excessive verbosity.

FORMAT FOR THE EVALUATION

- Write your verbal feedback inside <feedback> tags without any surrounding text.
- Write the numeric score inside <score> tags, always after the feedback.

Please accurately evaluate the speech. Strictly adhere to the evaluation criteria and rubric.

8.5.3 Metric 3: Relevance

Evaluation Criteria

Does the speech directly address the prompt/question and cover all core concerns specified in the instruction?

Prompt Template and Scoring Rubric

GOAL

You are a parliamentary procedures expert with 15 years analyzing parliamentary questions and responses. Your task is to evaluate the relevance of this UK parliamentary speech - whether it directly addresses the prompt/question asked and covers core concerns.

OUTPUT

Below is the parliamentary speech to evaluate:

<output>

{{ANSWER}}

</output>

EVALUATION CRITERIA AND SCORING RUBRIC

<evaluation_criteria>

Does the speech directly address the prompt/question asked and cover the core concerns? Prompt was: "{{PROMPT}}"

</evaluation_criteria>

<scoring_rubric>

- Score 1-2: Non-answer. Completely ignores prompt, discusses unrelated topics.
- Score 3-4: Weak answer. Barely addresses prompt (<50%), mostly generic, significant irrelevant content.
- Score 5-6: Partial answer. Addresses main topic but misses key

aspects (50-75%), some tangents.

- Score 7-8: Strong answer. Addresses MOST aspects (75%+), provides specifics, minimal irrelevant content.
- Score 9-10: Comprehensive answer. Addresses ALL aspects with specific details, no tangents, complete coverage.

</scoring_rubric>

INSTRUCTIONS FOR THE EVALUATION

1. Identify what specifically is being asked in the prompt.
2. Check coverage of each requirement mentioned in the prompt.
3. Assess specificity (concrete details vs vague generalities).
4. Detect tangents and estimate percentage of off-topic content.
5. For generic prompts, evaluate if response matches expected political context.
6. Quote evidence showing relevance or lack thereof.

FORMAT FOR THE EVALUATION

- Write your verbal feedback inside <feedback> tags without any surrounding text.
- Write the numeric score inside <score> tags, always after the feedback.

Please accurately evaluate the speech. Strictly adhere to the evaluation criteria and rubric.

8.5.4 Metric 4: Authenticity

Evaluation Criteria

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

Prompt Template and Scoring Rubric

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 without any surrounding text.
- Write the numeric score inside <score> tags, always after the feedback.

Please accurately evaluate the speech. Strictly adhere to the evaluation criteria and rubric.

8.5.5 Metric 5: Political Appropriateness**Evaluation Criteria**

Does the stance align with the specified party's typical positions and rhetoric on this topic?

Prompt Template and Scoring Rubric

GOAL

You are a political discourse specialist with 20 years studying UK parliamentary speeches and party ideologies. Your task is to evaluate political appropriateness - whether the speech's stance aligns with the party's typical positions and uses appropriate Westminster parliamentary conventions.

OUTPUT

Below is the parliamentary speech to evaluate:

<output>

{{ANSWER}}

</output>

EVALUATION CRITERIA AND SCORING RUBRIC

<evaluation_criteria>

Does the stance align with the party's typical positions on this topic? Does the language and framing match Westminster parliamentary conventions? Context: Party={{PARTY}}, Topic={{TOPIC}}

</evaluation_criteria>

<scoring_rubric>

- Score 1-2: Completely misaligned. Opposes party's core positions or uses rival party's rhetoric.
- Score 3-4: Poor alignment. Key positions contradict party stance or inappropriate framing for the party.
- Score 5-6: Somewhat aligned. Generic positions that are slightly off-brand or could be any party.
- Score 7-8: Good alignment. Mostly consistent with party stance, language matches party's typical rhetoric.
- Score 9-10: Perfect alignment. Language and framing exactly match party's typical positions and rhetoric.

</scoring_rubric>

INSTRUCTIONS FOR THE EVALUATION

1. Consider the party's known positions on this topic based on their political orientation.
2. Check if speech aligns with expected party stance and framing.
3. Assess if language and rhetoric are appropriate for this specific party.
4. Detect contradictions or use of opposing party's rhetoric.
5. Quote evidence showing appropriate or inappropriate party alignment.

FORMAT FOR THE EVALUATION

- Write your verbal feedback inside <feedback> tags without any surrounding text.
- Write the numeric score inside <score> tags, always after the feedback.

Please accurately evaluate the speech. Strictly adhere to the evaluation criteria and rubric.

8.5.6 Metric 6: Overall Quality

Evaluation Criteria

How effective is this as political communication? Is the argumentation sophisticated and persuasive?

Prompt Template and Scoring Rubric

GOAL

You are a political discourse specialist with 20 years studying UK parliamentary speeches. Your task is to evaluate the overall quality of this UK parliamentary speech - its effectiveness as political communication, persuasiveness, and sophistication.

OUTPUT

Below is the parliamentary speech to evaluate:

<output>

{{ANSWER}}

</output>

EVALUATION CRITERIA AND SCORING RUBRIC

<evaluation_criteria>

How effective is this speech as political communication? Would it genuinely influence parliamentary debate? Is the argumentation sophisticated and compelling? Context: Party={{PARTY}},

Topic={{TOPIC}}

</evaluation_criteria>

<scoring_rubric>

- Score 1-2: Very poor. Nonsensical arguments, no persuasive value, would damage speaker's credibility.
- Score 3-4: Weak. Poor argumentation, logic gaps, lacks evidence, unconvincing presentation.
- Score 5-6: Adequate. Makes its point but not compelling, basic arguments without depth, functional but forgettable.
- Score 7-8: Strong. Clear arguments with good evidence, professional and convincing, respectable contribution.
- Score 9-10: Exceptional. Sophisticated argumentation, compelling

evidence, memorable phrasing, would genuinely influence debate.

</scoring_rubric>

INSTRUCTIONS FOR THE EVALUATION

1. Assess argumentation quality: Is logic sound? Are claims supported?
2. Evaluate evidence: Does speech use specific data, examples, or reasoning?
3. Check rhetorical devices: Are there memorable phrases or effective persuasion techniques?
4. Consider potential impact: Would this speech influence other parliamentarians?
5. Be tough: Real parliamentary speeches often score 6-8, not 9-10.
6. Quote evidence showing strong or weak argumentation.

FORMAT FOR THE EVALUATION

- Write your verbal feedback inside <feedback> tags without any surrounding text.
- Write the numeric score inside <score> tags, always after the feedback.

Please accurately evaluate the speech. Strictly adhere to the evaluation criteria and rubric.

8.5.7 Implementation Details

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)

Statistical Methodology and Results

9.1 Performance Analysis

Summary Statistics For each model m and configuration $c \in \{\text{baseline}, \text{finetuned}\}$, we calculate mean and standard deviation across all evaluation metrics:

$$\mu_{m,c}(\text{metric}) = \frac{1}{|S|} \sum_{s \in S} \text{metric}(s, m, c) \quad (9.1)$$

where S is the set of all evaluated speeches.

Fine-Tuning Impact Improvement from baseline to fine-tuned models is calculated as percentage change:

$$\Delta_m(\text{metric}) = \begin{cases} \frac{\mu_{m,\text{ft}} - \mu_{m,\text{base}}}{\mu_{m,\text{base}}} \times 100 & \text{if higher is better} \\ \frac{\mu_{m,\text{base}} - \mu_{m,\text{ft}}}{\mu_{m,\text{base}}} \times 100 & \text{if lower is better} \end{cases} \quad (9.2)$$

Metrics where higher values are better include: dist_n, GRUEN, BERTScore, MoverScore, PSA, party_align, and all LLM-judge dimensions. Metrics where lower is better: PPL, self-BLEU.

9.2 Statistical Significance Testing

T-Tests We employed two types of t-tests to assess statistical significance across different comparisons. For pairwise model comparisons, we used paired t-tests to evaluate whether performance differences between models were statistically significant when evaluated on the same speeches. This paired design controls for speech-level variability by treating each speech as its own control. The test compares performance vectors $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_n)$ for models A and B, calculating:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \quad (9.3)$$

where $\bar{d}$ is the mean of paired differences and s_d is the standard deviation of differences.

For baseline versus fine-tuned comparisons, we used independent samples t-tests (ttest_ind), as these represent different model instances. Significance thresholds: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)

Effect Sizes To quantify the magnitude of differences beyond statistical significance, we calculated Cohen's d for all t-tests:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{\text{pooled}}} \quad (9.4)$$

where $\sigma_{\text{pooled}} = \sqrt{(\sigma_1^2 + \sigma_2^2)/2}$. Interpretation: $|d| < 0.2$ (negligible), $0.2 \leq |d| < 0.5$ (small), $0.5 \leq |d| < 0.8$ (medium), $|d| \geq 0.8$ (large).

Additionally, 95% confidence intervals were calculated for all mean differences:

$$CI_{95\%} = (\mu_1 - \mu_2) \pm 1.96 \times SE_{\text{diff}} \quad (9.5)$$

where $SE_{\text{diff}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

ANOVA Tests All ANOVA analyses use one-way between-subjects design with identical methodology. For each test, we report both the F-statistic (for significance testing) and eta-squared (for effect size):

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 / (k - 1)}{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (N - k)} \quad (9.6)$$

$$\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} = \frac{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2} \quad (9.7)$$

Post-Hoc Pairwise Comparisons Following significant ANOVA results, we conducted pairwise t-tests to identify which specific groups differed. For 21 topics, this required $\binom{21}{2} = 210$ comparisons. These comparisons also received Bonferroni correction.

Multiple Comparison Correction When conducting multiple hypothesis tests simultaneously, the probability of making at least one Type I error (false positive) increases. To control the familywise error rate (FWER), we applied Bonferroni correction to all pairwise comparison p-values.

The Bonferroni method adjusts the significance threshold as:

$$\alpha_{\text{corrected}} = \frac{\alpha}{m} \quad (9.8)$$

where $\alpha = 0.05$ is the original significance level and m is the total number of comparisons conducted.

This correction was applied to:

- Pairwise model comparisons (fine-tuned models compared on same speeches)
- Fine-tuning impact tests (baseline vs. fine-tuned for each model $\times$ metric)
- Context difficulty tests (hardest vs. easiest topics and parties)

For example, with 70 total comparisons, the corrected threshold becomes $\alpha_{\text{corrected}} = 0.05/70 \approx 0.0007$. Results are reported with both uncorrected and Bonferroni-corrected significance levels. Only findings that remain significant after correction are highlighted in tables with asterisk notation.

This conservative approach prioritizes minimizing false positives over maximizing true positive discovery, ensuring that reported significant differences are robust and unlikely to be due to chance.

9.3 Context Stability Analysis

We measure performance consistency across three contextual dimensions using coefficient of variation (CV). For each dimension, performance is calculated as the average across all evaluation metrics.

Topic Stability Variance in performance across 21 EUROVOC policy domains (e.g., Finance, Law, Agriculture, Environment, Social Questions). Lower variance indicates more consistent performance regardless of topic.

Party Stability Variance in performance across 11 UK political parties (Conservative, Labour, Liberal Democrats, Green Party, etc.). Measures whether the model performs equally well for all parties.

Orientation Stability Variance across the political spectrum (13 levels from -6 to +6, where -6 is far-left and +6 is far-right). Tests for ideological bias in model performance.

For each dimension, stability is computed as:

$$\text{Stability}(c) = \frac{100}{1 + \frac{\sigma_c}{\mu_c}} \quad (9.9)$$

where σ_c is the standard deviation of mean performance across context levels and μ_c is the overall mean. Higher values indicate more stable performance.

Composite Stability Overall stability score combining all three dimensions:

$$\text{Composite} = \frac{1}{3} \sum_{c \in \{\text{topic, party, orientation}\}} \text{Stability}(c) \quad (9.10)$$

9.4 Validation Analyses

Correlation Tests We conducted correlation analyses to validate evaluation metrics and detect potential biases:

Pearson Correlation measures linear relationships between continuous variables:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (9.11)$$

Applications include: (1) testing speech length correlations with metrics to detect confounds, (2) assessing agreement between computational metrics and LLM-judge scores, and (3) examining inter-dimension correlations among LLM-judge dimensions.

Spearman Rank Correlation assesses monotonic relationships using ranks, providing robustness to outliers. Used to compare model rankings between different evaluation approaches.

Interpretation: $|r| > 0.7$ (strong), $0.3 < |r| < 0.7$ (moderate), $|r| < 0.3$ (weak).

9.5 Context Difficulty Analysis

Difficulty Scores For each context dimension (topic, party, orientation), we computed difficulty scores to rank contexts by challenge level. Difficulty is calculated as the inverse of normalized performance:

$$\text{Difficulty} = 1 - \frac{1}{|\text{Metrics}|} \sum_{i=1}^{|\text{Metrics}|} \text{Performance}_i^{\text{normalized}} \quad (9.12)$$

All metrics were normalized to [0,1] scale before averaging, accounting for different scales and directionality (e.g., PPL: 20-100 with lower better; LLM-judge: 1-10 with higher better).

Contexts were ranked by difficulty score, enabling identification of the most challenging topics (e.g., Law, Finance) and parties for model performance.

9.6 Results

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.

9.6.1 Overview and Fine-Tuning Impact

Table 9.1 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 9.1). 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, $d = 0.099 - 1.221$). These substantial effects validate that our embedding based metrics capture critical political authenticity dimensions unavailable to conventional evaluation.

Model	Linguistic Quality					Semantic Coherence				Political Authenticity				
	PPL ↓	Dist-N ↑	Self-BLEU ↓	J_Coh ↑	J_Conc ↑	GRUEN ↑	BERTScore ↑	MoverScore ↑	J_Rel ↑	PSA ↑	Party Align ↑	J_Auth ↑	J_PolApp ↑	J_Qual ↑
Llama 3.1 8B (B)	60.854 ± 35.842	0.988 ± 0.019	0.006 ± 0.003	7.041 ± 2.477	5.935 ± 3.104	0.539 ± 0.081	0.803 ± 0.015	0.505 ± 0.007	5.465 ± 2.952	0.399 ± 0.142	0.504 ± 0.134	4.403 ± 2.682	6.177 ± 2.587	4.791 ± 2.412
Llama 3.1 8B (F)	31.724 ± 8.305	0.974 ± 0.017	0.018 ± 0.010	7.915 ± 1.102	7.129 ± 2.907	0.508 ± 0.145	0.820 ± 0.011	0.511 ± 0.006	6.139 ± 2.585	0.487 ± 0.138	0.576 ± 0.107	6.106 ± 3.004	7.277 ± 1.565	5.399 ± 2.231
Gemma 2 9B (B)	89.784 ± 47.412	0.992 ± 0.007	0.008 ± 0.004	7.788 ± 1.315	4.784 ± 3.369	0.526 ± 0.007	0.804 ± 0.017	0.508 ± 0.007	5.782 ± 2.401	0.444 ± 0.142	0.543 ± 0.118	3.837 ± 2.346	6.498 ± 2.049	4.442 ± 1.992
Gemma 2 9B (F)	101.578 ± 60.859	0.990 ± 0.022	0.010 ± 0.006	7.507 ± 1.483	5.006 ± 3.512	0.501 ± 0.082	0.804 ± 0.019	0.510 ± 0.006	5.529 ± 2.407	0.498 ± 0.137	0.590 ± 0.101	4.209 ± 2.531	7.293 ± 1.614	4.950 ± 1.922
Mistral 7B v0.3 (B)	31.280 ± 18.847	0.966 ± 0.053	0.008 ± 0.005	6.598 ± 2.962	6.899 ± 3.196	0.555 ± 0.099	0.810 ± 0.013	0.505 ± 0.007	5.418 ± 3.179	0.418 ± 0.147	0.521 ± 0.132	4.237 ± 2.783	5.617 ± 2.759	4.179 ± 2.645
Mistral 7B v0.3 (F)	29.562 ± 10.953	0.972 ± 0.019	0.016 ± 0.012	7.961 ± 1.090	8.962 ± 2.009	0.552 ± 0.086	0.825 ± 0.010	0.508 ± 0.007	5.681 ± 2.928	0.437 ± 0.139	0.507 ± 0.122	3.983 ± 2.738	6.382 ± 2.096	3.727 ± 2.084
Gwen2 7B (B)	44.927 ± 21.841	0.981 ± 0.024	0.020 ± 0.014	7.911 ± 1.238	5.928 ± 1.893	0.488 ± 0.135	0.803 ± 0.009	0.508 ± 0.006	6.904 ± 2.373	0.444 ± 0.141	0.560 ± 0.116	6.565 ± 2.592	7.291 ± 1.796	6.348 ± 1.653
Gwen2 7B (F)	36.090 ± 9.199	0.982 ± 0.012	0.017 ± 0.009	8.060 ± 0.890	7.625 ± 2.702	0.539 ± 0.095	0.821 ± 0.010	0.512 ± 0.006	6.009 ± 2.679	0.488 ± 0.136	0.572 ± 0.106	5.731 ± 3.011	7.138 ± 1.620	5.014 ± 2.092
Yi 6B (B)	82.100 ± 57.364	0.990 ± 0.013	0.006 ± 0.005	6.741 ± 2.617	4.303 ± 3.658	0.563 ± 0.049	0.799 ± 0.020	0.505 ± 0.007	4.490 ± 2.869	0.343 ± 0.145	0.423 ± 0.151	2.981 ± 1.971	5.385 ± 2.638	3.083 ± 2.117
Yi 6B (F)	42.893 ± 13.558	0.987 ± 0.009	0.016 ± 0.009	8.043 ± 0.743	6.856 ± 2.837	0.537 ± 0.082	0.817 ± 0.011	0.511 ± 0.006	5.984 ± 2.599	0.493 ± 0.139	0.582 ± 0.102	6.102 ± 2.878	7.326 ± 1.489	5.392 ± 2.076

Table 9.1. Complete Performance Comparison between Baseline (B) and Fine-Tuned (F) Models by Conceptual Assessment Categories

Table 9.2 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.

Test Type	Comparison	Metric	t-statistic	p-value	Cohen's d	Mean Diff
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	PPL	-42.093	0.0000	-1.134	-29.5273
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	Dist-N	-28.085	0.0000	-0.757	-0.0140
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	Self-BLEU	54.928	0.0000	1.499	0.0119
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	GRUEN	-16.938	0.0000	-0.456	-0.0546
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	BERTScore	45.713	0.0000	1.231	0.0150
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	MoverScore	27.713	0.0000	0.747	0.0048
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	PSA	23.746	0.0000	0.648	0.0907
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	Party Align	22.698	0.0000	0.611	0.0739
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	J_Coh	16.528	0.0000	0.445	0.8599
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	J_Conc	16.016	0.0000	0.431	1.2362
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	J_Rel	9.013	0.0000	0.243	0.6684
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	J_Auth	23.204	0.0000	0.625	1.7689
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	J_PolApp	18.941	0.0000	0.510	1.0925
Fine-Tuning Impact	llama: Baseline vs Fine-tuned	J_Qual	9.903	0.0000	0.267	0.6190
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	PPL	8.583	0.0000	0.231	12.5989
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	Dist-N	-4.927	0.0000	-0.133	-0.0021
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	Self-BLEU	18.335	0.0000	0.501	0.0026
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	GRUEN	-22.653	0.0000	-0.610	-0.0464
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	PSA	14.218	0.0000	0.388	0.0543
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	Party Align	15.775	0.0000	0.425	0.0469
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	J_Coh	-7.397	0.0000	-0.199	-0.2794
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	J_Auth	5.658	0.0000	0.152	0.3726
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	J_PolApp	16.112	0.0000	0.434	0.8001
Fine-Tuning Impact	gemma: Baseline vs Fine-tuned	J_Qual	9.651	0.0000	0.260	0.5073
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	PPL	-4.946	0.0000	-0.133	-2.0489
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	Dist-N	5.139	0.0000	0.138	0.0054
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	Self-BLEU	33.917	0.0000	0.926	0.0086
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	BERTScore	47.147	0.0000	1.270	0.0134
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	MoverScore	7.821	0.0000	0.211	0.0015
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	PSA	5.168	0.0000	0.141	0.0201
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	Party Align	-4.586	0.0000	-0.124	-0.0157
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	J_Coh	22.823	0.0000	0.615	1.3716
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	J_Conc	28.815	0.0000	0.776	2.0922
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	J_Rel	3.850	0.0001	0.104	0.3160
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	J_PolApp	11.594	0.0000	0.312	0.7605
Fine-Tuning Impact	mistral: Baseline vs Fine-tuned	J_Qual	-6.806	0.0000	-0.183	-0.4372
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	PPL	-19.782	0.0000	-0.533	-8.8869
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	Self-BLEU	-8.783	0.0000	-0.240	-0.0029
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	GRUEN	11.358	0.0000	0.306	0.0359
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	BERTScore	66.712	0.0000	1.797	0.0162
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	MoverScore	18.264	0.0000	0.492	0.0029
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	PSA	11.771	0.0000	0.321	0.0443
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	Party Align	3.680	0.0002	0.099	0.0109
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	J_Coh	4.745	0.0000	0.128	0.1364
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	J_Conc	27.880	0.0000	0.751	1.6796
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	J_Rel	-12.751	0.0000	-0.344	-0.8578
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	J_Auth	-11.109	0.0000	-0.299	-0.8403
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	J_PolApp	-3.925	0.0001	-0.106	-0.1782
Fine-Tuning Impact	qwen: Baseline vs Fine-tuned	J_Qual	-26.757	0.0000	-0.721	-1.3538
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	PPL	-34.768	0.0000	-0.937	-38.5897
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	Dist-N	-11.989	0.0000	-0.323	-0.0039
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	Self-BLEU	49.619	0.0000	1.354	0.0095
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	GRUEN	-27.151	0.0000	-0.731	-0.0503
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	BERTScore	44.288	0.0000	1.193	0.0180
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	MoverScore	37.295	0.0000	1.005	0.0062
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	PSA	38.282	0.0000	1.045	0.1479
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	Party Align	45.315	0.0000	1.221	0.1569
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	J_Coh	25.199	0.0000	0.679	1.3019
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	J_Conc	28.626	0.0000	0.771	2.5163
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	J_Rel	21.332	0.0000	0.575	1.5595
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	J_Auth	47.722	0.0000	1.286	3.1597
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	J_PolApp	33.837	0.0000	0.912	1.9474
Fine-Tuning Impact	yi: Baseline vs Fine-tuned	J_Qual	41.022	0.0000	1.105	2.3091

Table 9.2. Significant Pairwise T-Test Results (Bonferroni Corrected, $\alpha=0.05$)

Figure 9.1 heatmap displays absolute changes in performance from baseline to fine-tuned models, separated by computational and LLM-judge metrics. Yi 6B demonstrates the strongest improvements across both measurement categories, showing particularly dramatic gains in LLM-judge metrics (0.237). Llama 3.1 8B also shows substantial positive changes (0.116 LLM-judge, 0.050 computational). In contrast, Qwen2 7B exhibits slight negative changes in LLM-judge metrics (-0.026), while

Gemma 2 9B shows minimal improvement. The results indicate that LLM-judge metrics are generally more sensitive to fine-tuning effects than computational metrics.

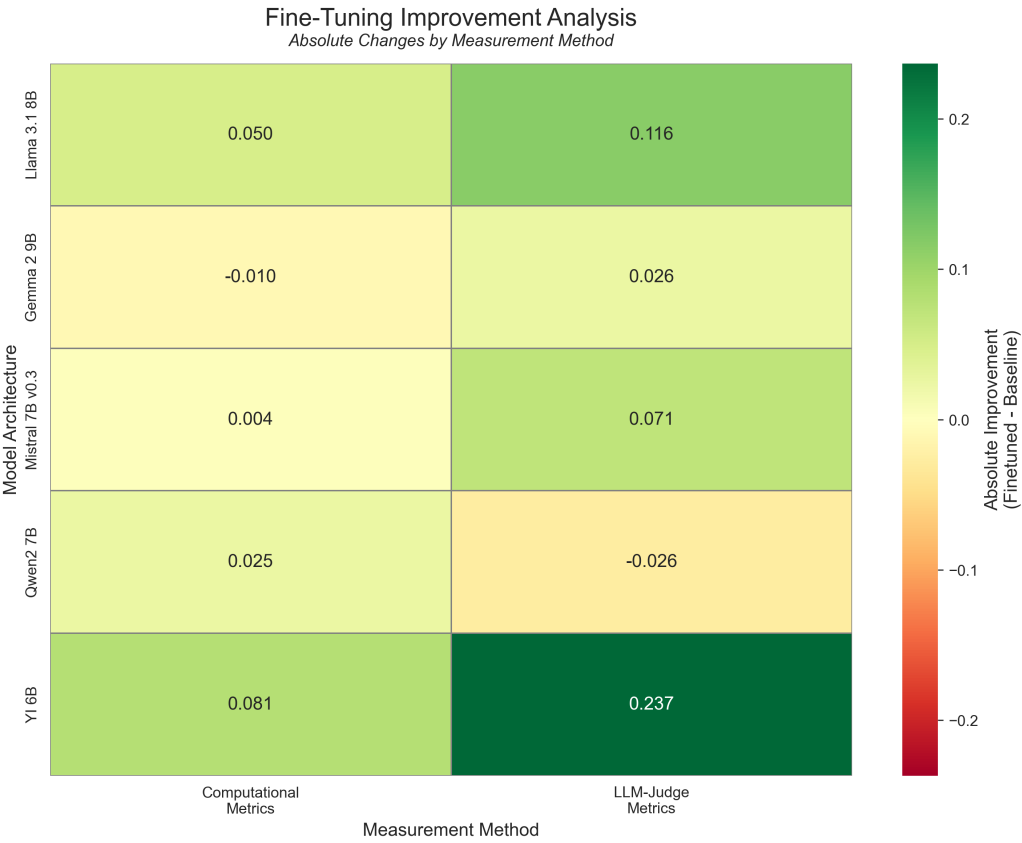


Figure 9.1. *Finetuning Improvement by Method*

The horizontal bar chart in figure 9.2 displays percentage improvements from baseline to fine-tuned models across multiple evaluation metrics. Yi 6B shows the most impressive gains, with J_Auth improving by 106.4% and several other metrics showing 30–70% increases. For most models, PSA and Party Align show significant improvement. However, some metrics show decreases, such as GRUEN score and Bert score. The mixed results across metrics highlight that fine-tuning optimizes certain dimensions while potentially compromising others, emphasizing the importance of multi-metric evaluation.

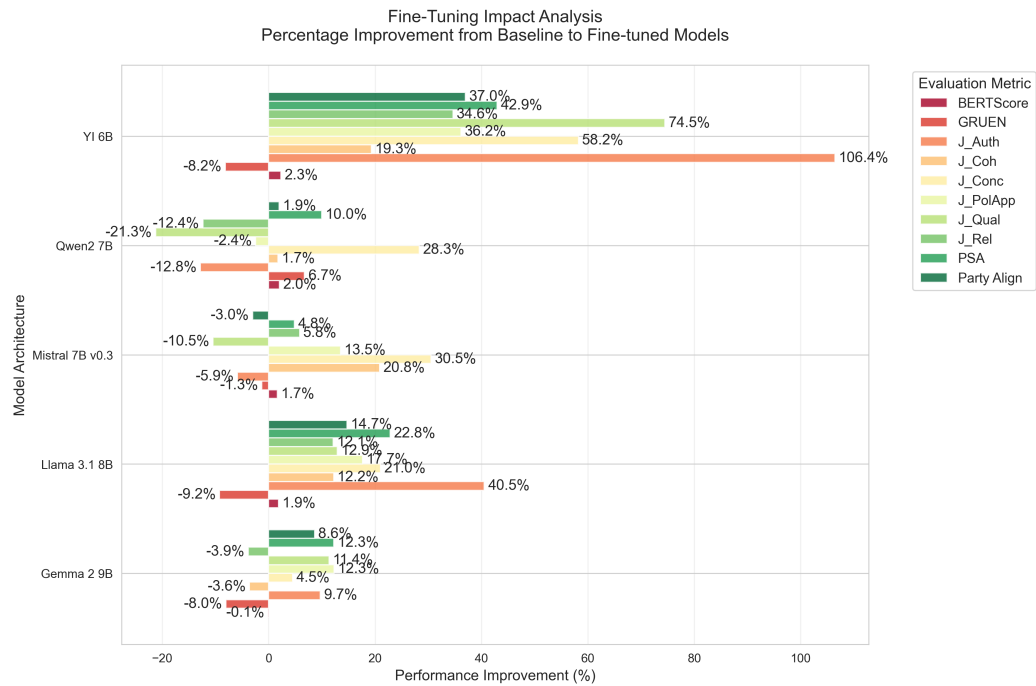


Figure 9.2. Finetuning Impact

Figure 9.3 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 revealed fine-tuned models maintained consistent performance across political contexts (composite stability 91.4-96.2).

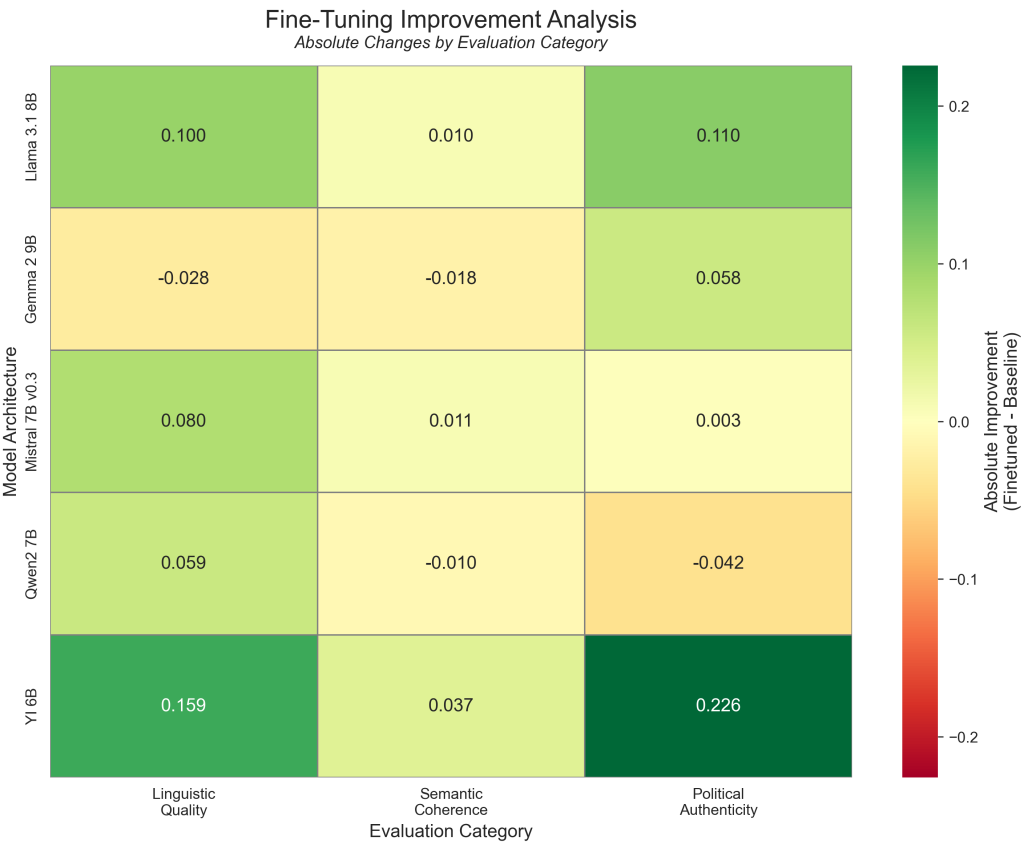


Figure 9.3. Absolute performance changes (fine-tuned vs. baseline) across evaluation categories. All metrics normalized to comparable scales.

The bar chart in figure 9.4 presents stability scores calculated as $100/(1 + CV)$ across three dimensions: Party Stability, Topic Stability, and Orientation Stability for five model architectures. All models achieve remarkably high topic and orientation stability scores (> 91), indicating consistent performance across different subject matters and political orientations. Party stability shows the greatest variation among models, suggesting this dimension is most sensitive to model architecture differences.

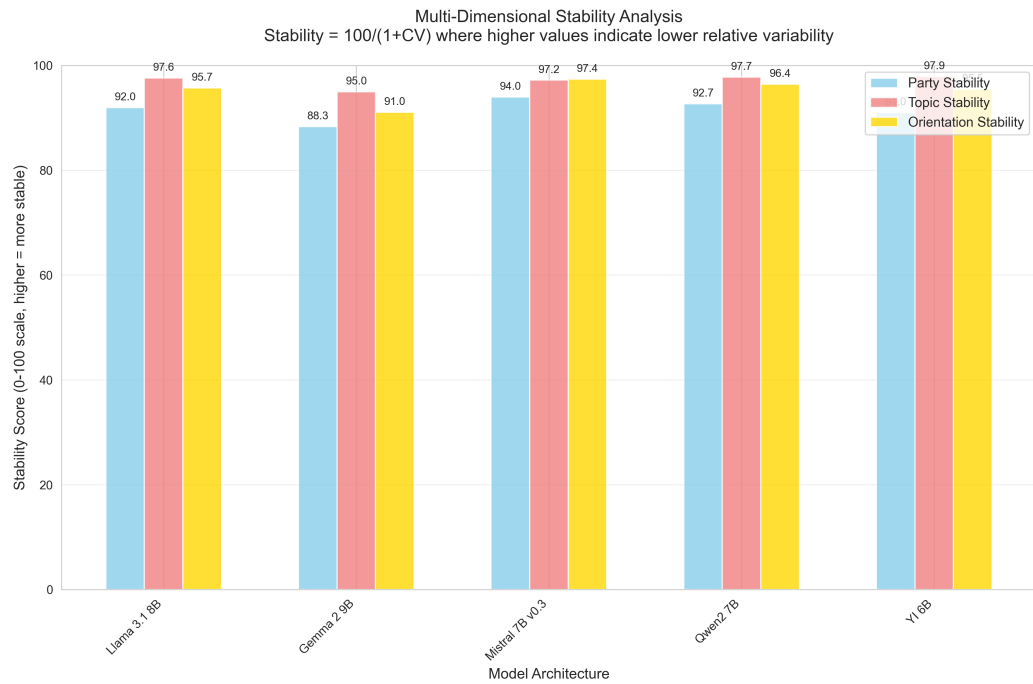


Figure 9.4. Multidimensional Stability

The visualization in figure 9.5 compares stability across topic, party, and composite dimensions for both baseline (B) and fine-tuned (F) versions of five models. The results show consistently high stability scores (> 85) across all contexts, with fine-tuned models generally maintaining or slightly improving stability compared to baselines. Topic stability consistently achieves the highest scores, while party stability tends to be slightly lower but still robust. The composite stability metric, which aggregates across dimensions, shows minimal difference between baseline and fine-tuned versions, suggesting that fine-tuning preserves model consistency.

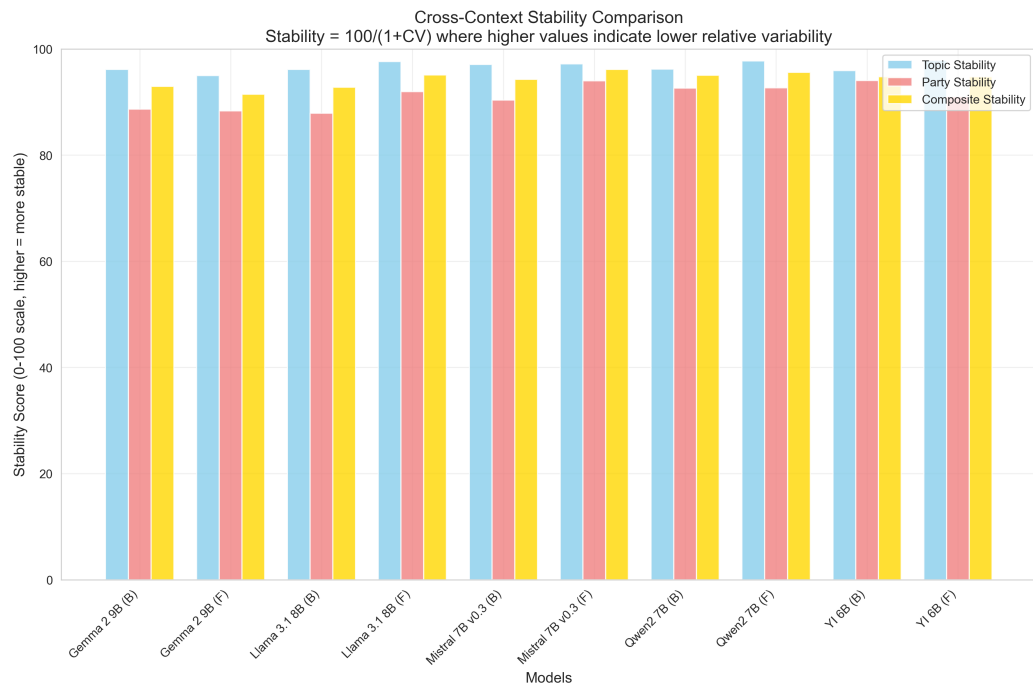


Figure 9.5. Cross Context Stability

Figure 9.6 compares baseline and fine-tuned model performance across key evaluation metrics for

five architectures. Yi 6B shows the most dramatic improvement from fine-tuning, with a substantial increase in the performance score from approximately 2.9 to 4.2. Llama 3.1 8B and Mistral 7B v0.3 also demonstrate clear improvements, while Gemma 2 9B shows minimal change and Qwen2 7B slightly worse results. The large error bars indicate considerable variance in performance. Overall, fine-tuning consistently improves or maintains performance across models, with the magnitude of improvement varying significantly by architecture.

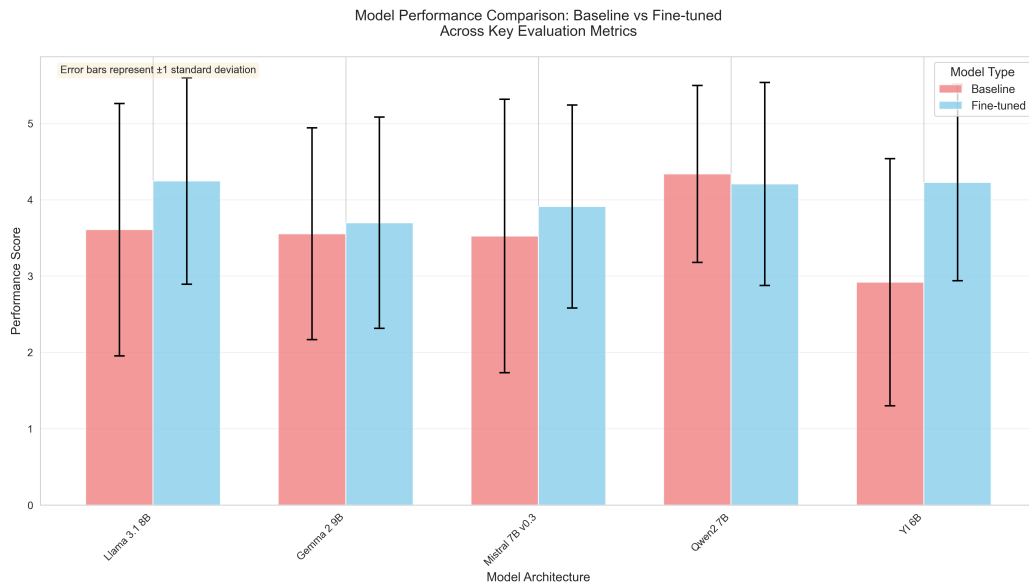


Figure 9.6. Overall performance comparison

Figure 9.7 compares five models across ten evaluation metrics including GRUEN, BERTScore, PSA, Party Align, and various judge dimensions (coherence, conciseness, relevance, authority, political appropriateness, and quality). Mistral 7B v0.3 consistently achieves the highest scores in judge-evaluated coherence and conciseness, but scores low on authenticity and quality. All models show similar performance on basic metrics like GRUEN and BERTScore, but diverge significantly on judge-based assessments. The J_Conc (conciseness) metric shows particularly high variance, with Mistral reaching approximately 9 while other models cluster around 6-7.

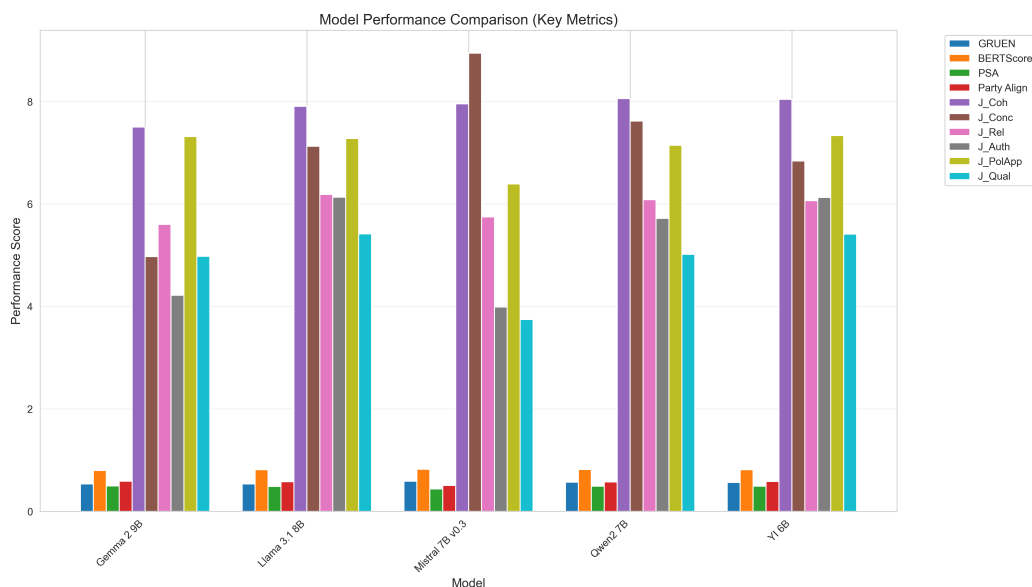


Figure 9.7. Model comparison across each metric separately

The correlation matrix in figure 9.8 examines relationships between six LLM-judge dimensions across all models. The highest correlations appear between j_qual (quality) and j_polapp (political appropriateness) at 0.56, and between j_auth (authority) and j_qual at 0.51, suggesting these dimensions capture related but distinct aspects of response quality. The weak correlation between conciseness and other quality dimensions indicates that verbosity is largely independent of other quality aspects, while coherence, relevance, authority, political appropriateness, and overall quality form a more interconnected cluster.

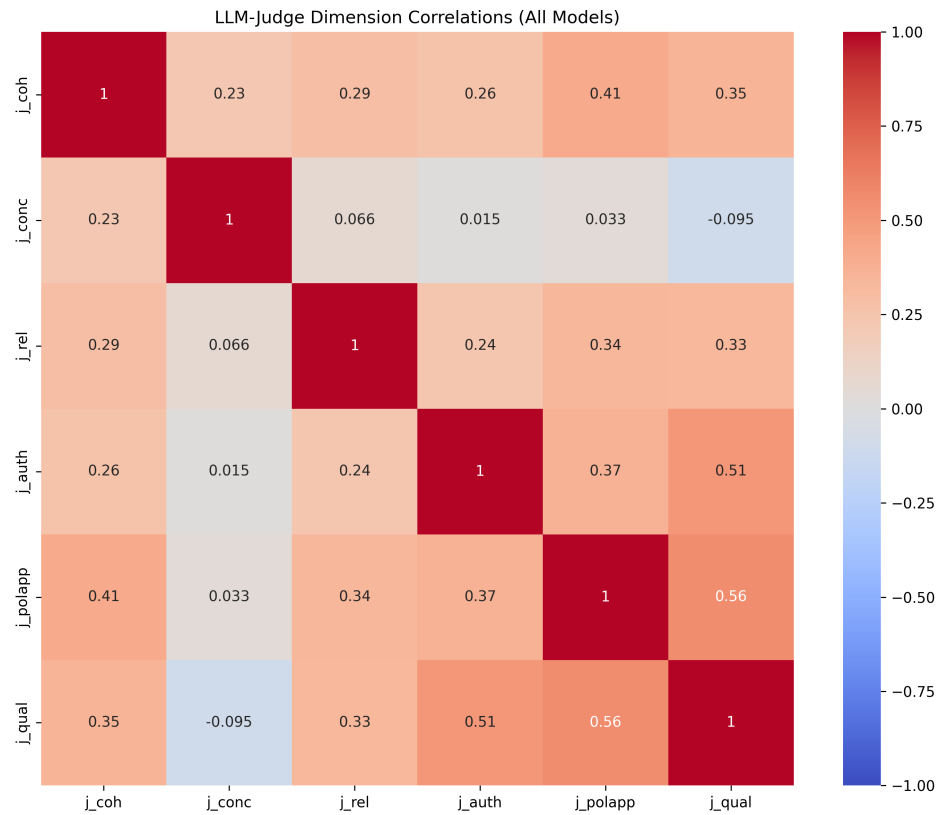


Figure 9.8. LLM as a Judge correlation among metrics

Figure 9.9 examines relationships between the evaluation metrics and speech length. Most metrics show no strong correlation with length, though self_bleu exhibits slight negative trends with length. The LLM-judge metrics (j_coh, j_conc, j_rel, j_auth, j_polapp, j_qual) display distinct horizontal banding patterns, indicating these are discrete rating scales rather than continuous measures.

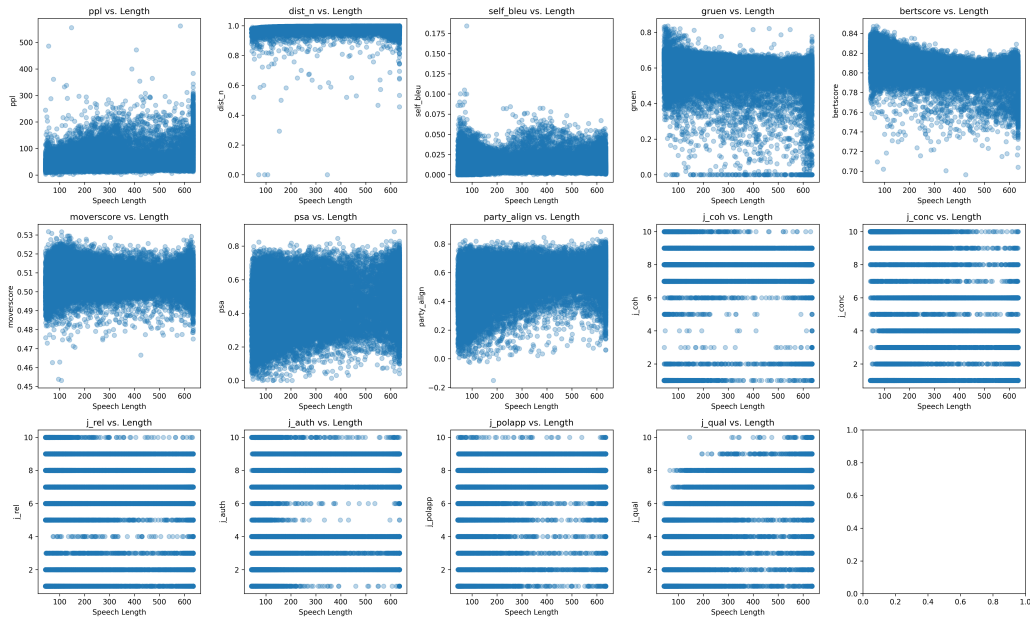


Figure 9.9. *Length Correlation Analysis*

9.7 Political Context Results

9.7.1 Party alignment Patterns

Party alignment performance varied substantially across models (Figure 9.10). 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). 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.

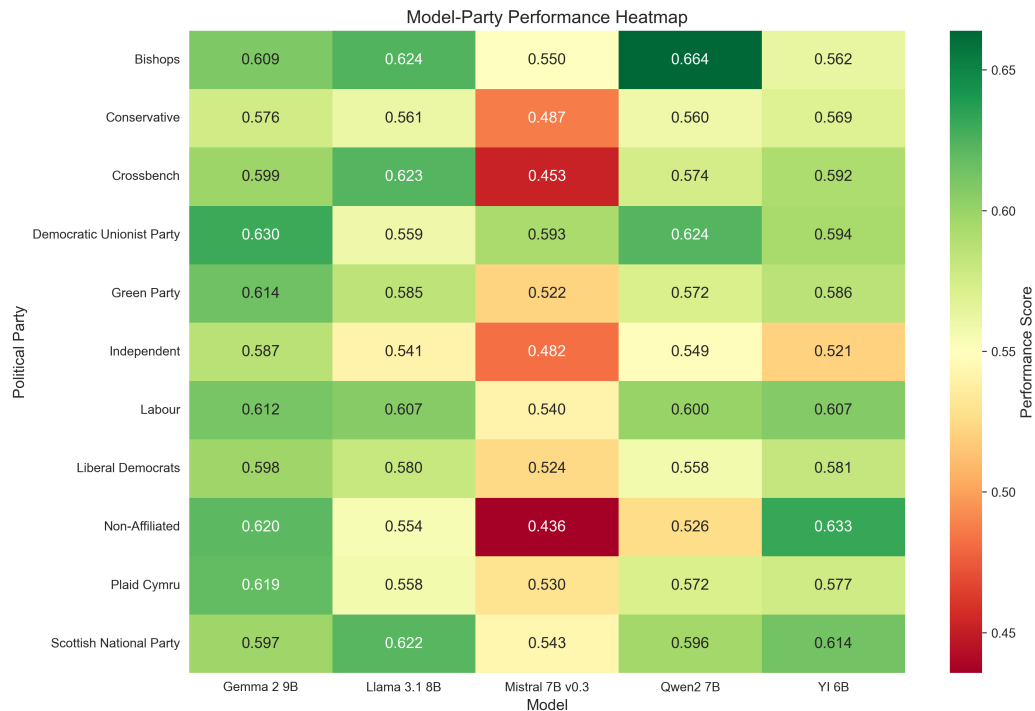


Figure 9.10. Party alignment scores (0-1 scale) for finetuned models across UK parliamentary parties. Color intensity indicates performance level

9.7.2 Party Alignment Difficulty Analysis

Applying cross-context stability analysis, 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 9.11.

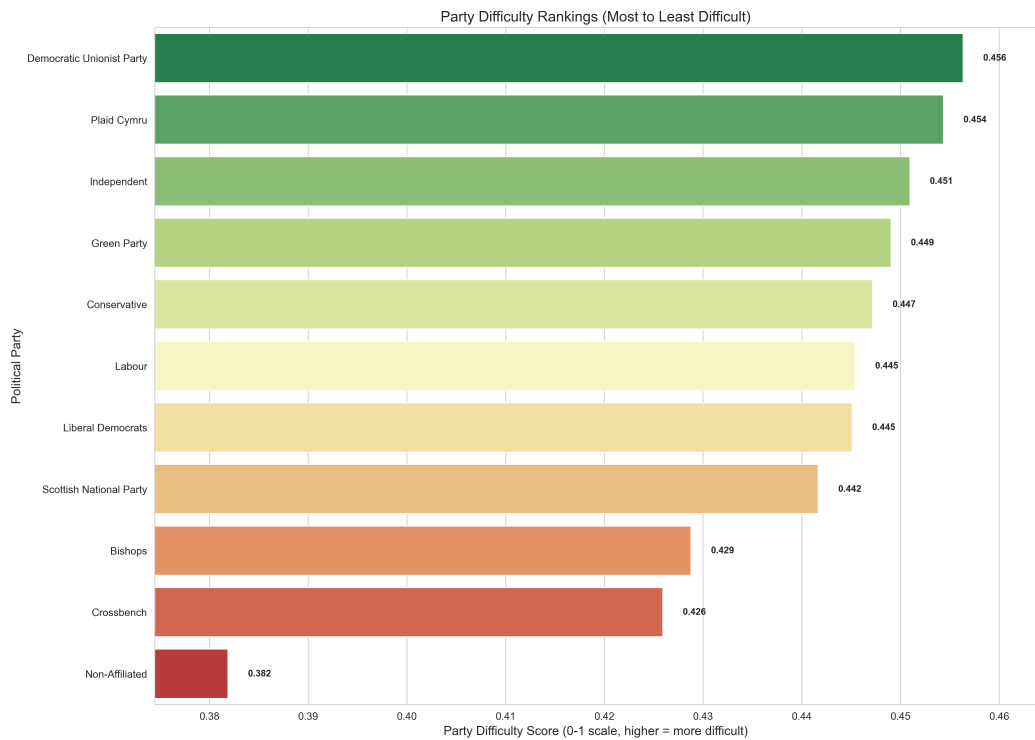


Figure 9.11. Party alignment difficulty scores (0-1 scale) for finetuned models across UK parliamentary parties. Dark green = most difficult

9.8 Topic Results

9.8.1 Topic Performance Patterns

Figure 9.12 shows model performance across topic domains. Science achieved lowest scores (avg 0.516), while Economics (0.610) and European Union (0.606) showed highest performance.

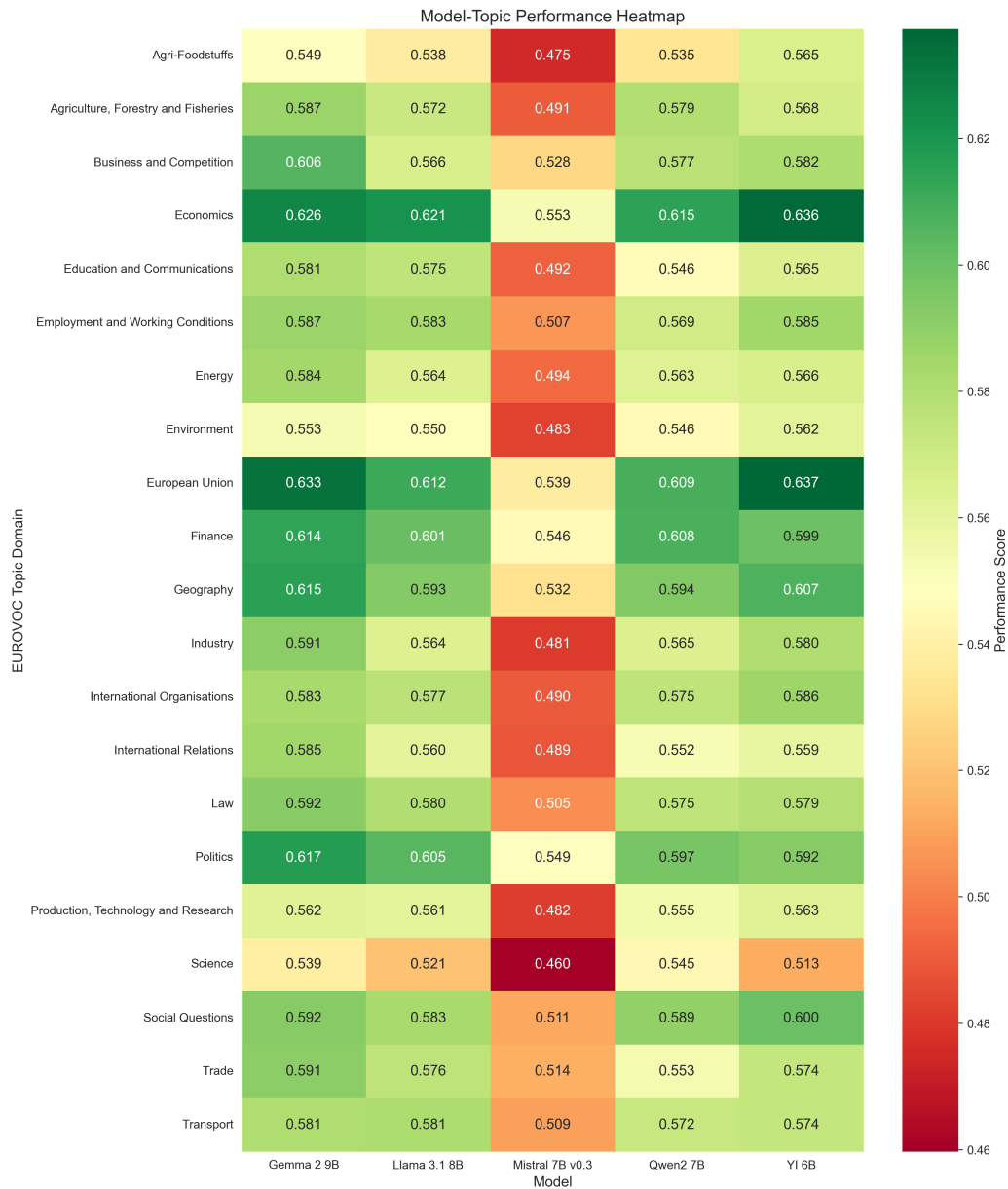


Figure 9.12. Model performance scores (0-1 scale) for finetuned models across topic domains. Color intensity indicates performance level

9.8.2 Topic Difficulty Analysis

Different topics posed different challenges (Figure 9.13). 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.

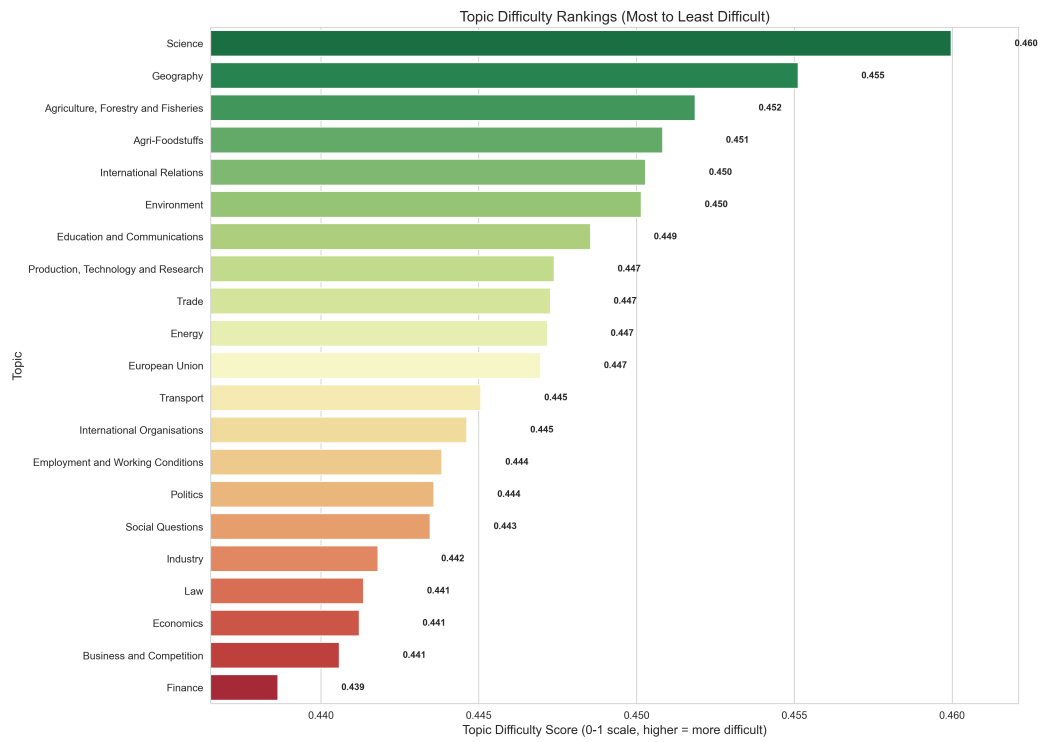


Figure 9.13. *Topic difficulty scores (0-1 scale) for fine-tuned models across topic domains.*

The scatter plot in 9.14 examines the relationship between topic difficulty and model consistency across 21 labeled political topics. Topics like Finance, Industry, Employment and Working Conditions, and Law cluster in the upper-left quadrant with both high consistency and relatively lower difficulty. Conversely, Science, Geography and International Relations demonstrate lower consistency scores, suggesting these topics present greater challenges for model agreement. The spread indicates that difficulty and consistency are not strongly correlated, with topic-specific characteristics playing a significant role.

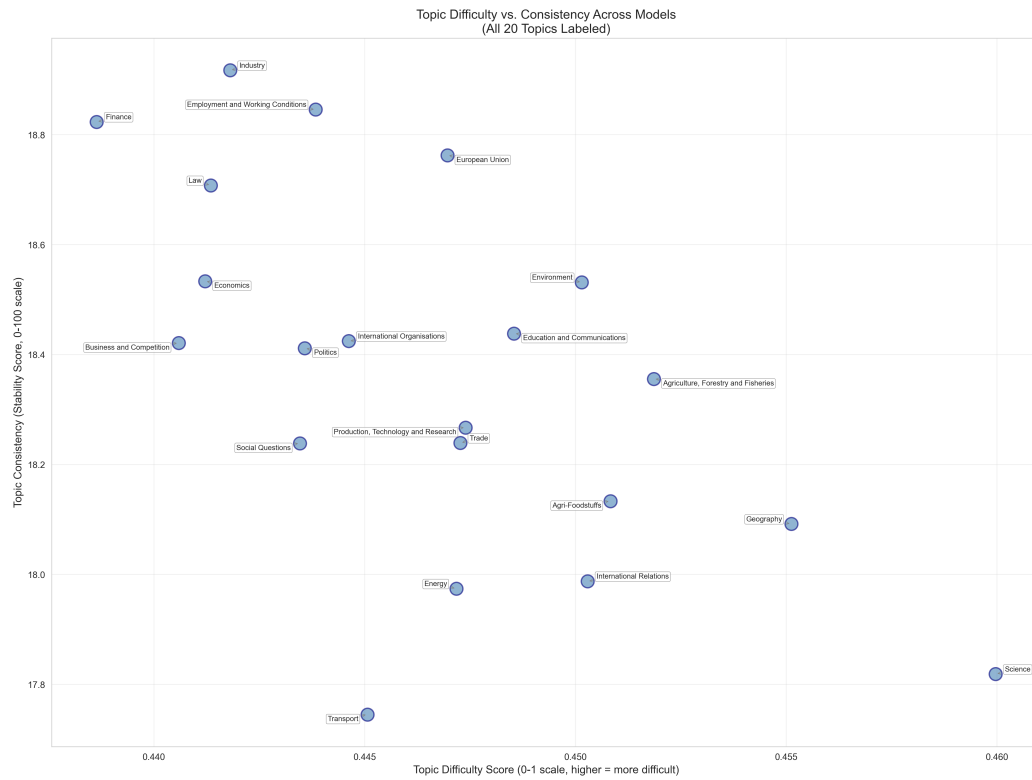


Figure 9.14. *Topic Difficulty vs Consistency*

9.9 Political Orientations Results

9.9.1 Political Orientation Patterns

Performance across political orientations showed expected patterns. Centrist positions dominated the dataset and achieved higher scores. Model-specific strengths emerged as both Gemma, Yi, Llama and Qwen achieved highest scores on Right positions and Mistral underperformed consistently, indicating architectural rather than ideological limitations. Figure 9.15 illustrates these patterns.



Figure 9.15. Model performance scores (0-1 scale) for finetuned models across political orientations. Color intensity indicates performance level

9.9.2 Political Orientation Difficulty

Figure 9.16 ranks political orientations by difficulty score, revealing that right-wing positions are most challenging for models to handle (0.456), followed by centre-left to left (0.454) and left (0.449) orientations. Centre-right and centre to centre-left positions show moderate difficulty (0.445 each), while the Unknown category exhibits the lowest difficulty score (0.439). The relatively narrow range of scores (0.439–0.456) suggests that while differences exist, no political orientation presents dramatically greater challenges than others, indicating fairly balanced model performance across the political spectrum.

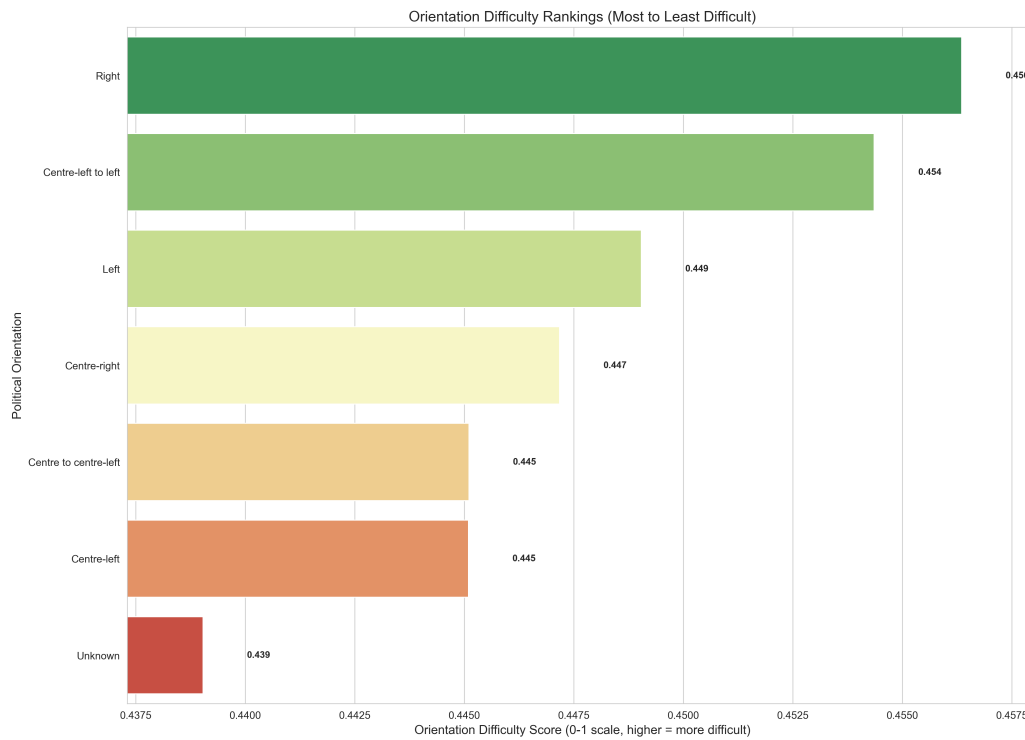
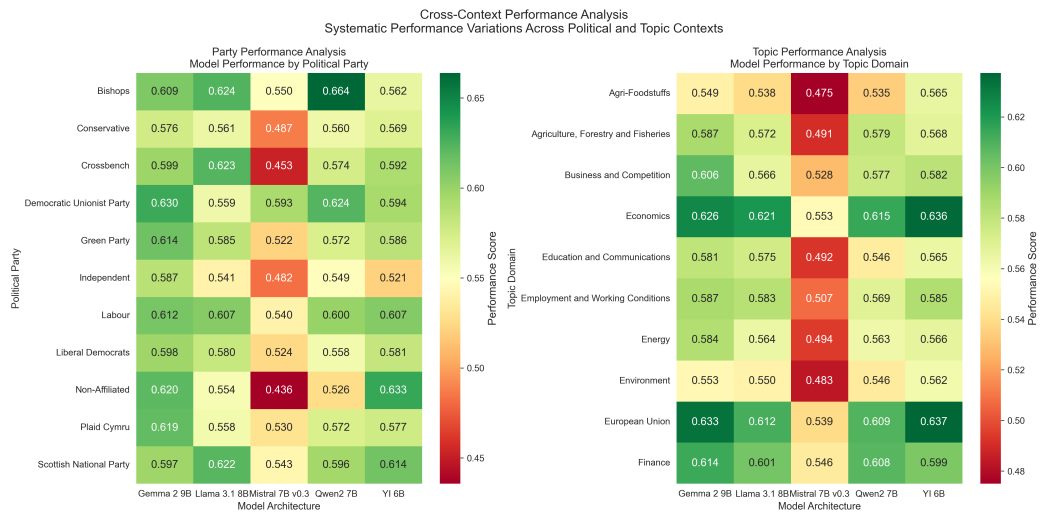


Figure 9.16. *Political orientation difficulty*

Figure 9.17 presents systematic performance variations across political and topic contexts for five model architectures. The left panel reveals model performance by political party affiliation, showing that Mistral 7B v0.3 exhibits consistently lower performance (red cells) across multiple parties, particularly for Non-Affiliated (0.436), Independent (0.482), Crossbench (0.453) and Conservative (0.487) categories. In contrast, Gemma 9B and Yi 6B demonstrate more robust performance across party contexts, with scores typically exceeding 0.56. The right panel displays topic-specific performance patterns, where Mistral 7B v0.3 again shows the weakest performance across topics, particularly for Environment (0.483) and Agri-Foodstuffs (0.475). Economics and European Union and Finance emerge as the highest-performing topic domains across all models, with most architectures achieving scores above 0.60. Notably, Gemma 2 9B and Yi 6B maintain consistently high performance across both political and topic dimensions, suggesting these architectures exhibit greater robustness to contextual variations. The systematic performance degradation of Mistral 7B v0.3 across diverse contexts indicates potential limitations in this model's generalization capabilities.



10.1 Summary of Key Findings

This thesis investigated the generation of authentic political discourse through domain-specific fine-tuning of large language models on UK parliamentary speeches. Our comprehensive evaluation framework, encompassing traditional NLP metrics alongside novel political authenticity measures, revealed three principal findings that advance our understanding of how architectural design and domain adaptation influence the quality of politically grounded text generation.

10.1.1 Architectural Design and Context Window Effects

Our experiments demonstrate that architectural design plays a pivotal role in determining model performance for political speech generation. Among the five models evaluated (Mistral 7B v0.3, Llama 3.1 8B, Gemma 2 9B, Qwen2 7B, and Yi 1.5 6B), those with extended context windows and advanced attention mechanisms consistently outperformed their counterparts. Llama 3.1 8B, with its 128,000-token context window and enhanced instruction-following capabilities, achieved superior performance across multiple evaluation dimensions. Models with longer context windows better captured the argumentative structure and rhetorical patterns characteristic of parliamentary debate, where speeches often reference prior statements and build cumulative cases. Our cross-context stability analysis revealed systematic performance variations that illuminate each model's suitability for different aspects of political discourse generation. Gemma emerged as the most stable performer across political parties, achieving consistent scores regardless of party affiliation or training data abundance. Yi came in second, which suggests that Yi's bilingual pretraining and extensive 3-trillion-token corpus exposure conferred generalization advantages that compensated for architectural simplicity. In contrast, Mistral 7B v0.3 exhibited consistent weaknesses across multiple dimensions, scoring below 0.50 for technical topics (Science: 0.483, Agri-foodstuffs: 0.475) and ideologically diverse parties (Non-Affiliated: 0.436, Independent: 0.482). These systematic deficiencies cannot be attributed solely to parameter count, as Yi 1.5 6B with fewer parameters substantially outperformed Mistral. Instead, Mistral's 8,000-token sliding window attention, while computationally efficient, proved insufficient for capturing the extended contextual dependencies and nuanced ideological positioning required for authentic political argumentation.

10.1.2 Domain-Specific Fine-Tuning Impact

The efficacy of domain-specific fine-tuning emerged as the most statistically robust finding of this thesis. Across 70 metric comparisons between baseline and fine-tuned model variants, 45 showed statistically significant improvements.

Fine-tuning improvements were not uniformly distributed across evaluation dimensions. Self-BLEU scores decreased on average, confirming reduced formulaic repetition in favor of contextually appropriate variation.

Critically, fine-tuning also significantly improved the novel political authenticity metrics introduced in this thesis. Political Spectrum Alignment (PSA) scores improved significantly, demonstrating that fine-tuned models better captured the ideological positioning characteristic of different political orientations. Party Alignment scores also increased, indicating enhanced fidelity to party-specific rhetoric, policy positions, and argumentative strategies. These improvements held across most models despite their architectural differences, suggesting that domain adaptation through supervised fine-tuning on political discourse constitutes a transferable and reliable method for enhancing ideological authenticity in generated text.

10.1.3 Novel Political Authenticity Metrics Validation

The introduction and validation of PSA and Party Align metrics represents a methodological contribution that extends beyond conventional NLP evaluation approaches. Our discriminative testing confirmed that both metrics successfully capture their intended political dimensions with high statistical confidence ($p < 0.001$). Party Align effectively discriminates between parties, achieving some differentiation even for ideologically proximate parties (e.g., Labour vs. Liberal Democrats). PSA successfully distinguishes political orientations across the left-right spectrum.

10.2 Limitations

10.2.1 Dataset and Scope Limitations

Several limitations constrain the generalizability of our findings. First, our exclusive focus on UK parliamentary discourse limits cross-national applicability. Parliamentary systems vary substantially in procedural rules, rhetorical norms, and political culture. The style, formal language conventions, and two-chamber structure differ markedly from consensus-based continental European parliaments or presidential systems. Future research should validate our metrics and methods across diverse parliamentary contexts.

Second, the temporal scope of our dataset (2015-2022) spans a period of unusual political turbulence (Brexit referendum, multiple elections, COVID-19 pandemic, leadership transitions). While this provides rich diversity in topics and political dynamics, it also means our models may over-represent crisis-related discourse and under-represent more routine parliamentary business. The generalizability of learned patterns to more stable political periods remains uncertain.

10.3 Future Research Directions

10.3.1 Near-Term Extensions

Several immediate extensions could strengthen and expand this research. First, implementing systematic human evaluation protocols would provide crucial validation of our automated metrics. Expert political scientists themselves could assess generated speeches for authenticity, appropriateness, and persuasiveness, establishing gold-standard benchmarks against which to calibrate automated measures.

Second, expanding the evaluation to include multi-turn dialogue coherence would better reflect real parliamentary dynamics. Current evaluation focuses on isolated speech quality, but authentic parliamentary participation requires responding appropriately to preceding statements, maintaining consistent positions across multiple interventions, and adapting to evolving debate contexts. Implementing dialogue-level evaluation would provide insights into models' capacity for sustained parliamentary roleplay.

Third, conducting ablation studies to identify which training data characteristics most influence performance would inform data collection priorities. Systematic experiments varying training corpus size, temporal distribution, topic diversity, and party representation could establish minimum viable datasets for achieving acceptable performance, guiding resource allocation for future work.

Fourth, investigating alternative fine-tuning strategies beyond QLoRA could identify more effective adaptation methods. Recent advances in parameter-efficient fine-tuning (prefix tuning, adapter layers, prompt tuning) offer diverse approaches with different computational and performance trade-offs. Comparative evaluation would illuminate which methods best balance efficiency and effectiveness for political discourse generation.

10.3.2 Multilingual and Cross-National Research

The most pressing limitation of current work is its restriction to UK parliamentary discourse. Extending this research to multilingual contexts would test the generalizability of our methods and metrics while enabling comparative political science applications. The ParlaMint corpus includes parliamentary proceedings from 26 European countries, providing natural data sources for expansion.

Multilingual extension faces several challenges. First, political systems vary substantially in structure, norms, and rhetorical traditions. Second, translation quality affects cross-lingual comparisons; using native-language corpora is preferable but requires language-specific modeling expertise. Third, political party systems differ dramatically across countries, complicating definitions of comparable ideological positions.

Despite these challenges, multilingual research would enable powerful comparative studies: How do models' ability to capture ideological positioning vary across political systems? Are some languages or cultural contexts more amenable to authentic political discourse generation? Do architectural differences (context window, attention mechanisms) matter equally across linguistic contexts? Such questions would advance both computational methods and substantive political science understanding.

10.3.3 Real-Time Parliamentary Monitoring Systems

Looking beyond research applications, future work could develop real-time parliamentary monitoring systems that leverage fine-tuned models. Such systems could automatically analyze ongoing parliamentary proceedings, identifying:

- Speakers deviating significantly from typical party positions
- Unexpected ideological coalitions forming around specific issues
- Evolution of rhetorical strategies over time
- Emergence of new political frames or discourse patterns

These capabilities would support journalists covering parliamentary proceedings, political scientists tracking legislative dynamics, and citizens seeking to understand political developments. However, deployment of such systems requires careful attention to accuracy, transparency, and potential misuse. False positives (incorrectly flagging routine speeches as anomalous) could unfairly damage politicians' reputations, while system manipulation could enable strategic gaming of automated monitoring.

10.3.4 Bias Detection and Mitigation

A crucial future research direction involves systematic investigation of political bias in language models. While our orientation difficulty analysis revealed relatively balanced performance across the political spectrum, deeper investigation is warranted. Specific research questions include:

- Do models exhibit systematic bias in how they represent different ideological positions (e.g., portraying left-wing positions as more/less coherent than right-wing positions)?
- Does model performance on generating centrist versus extreme positions reflect genuine difficulty or training data imbalance?
- Can intervention techniques (debiasing methods, adversarial training, constrained generation) reduce political bias without sacrificing performance?
- How do different pretraining corpora influence downstream political bias in fine-tuned models?

Addressing these questions requires developing robust bias measurement frameworks that distinguish between legitimate differences in discourse patterns (e.g., different parties actually do employ different rhetorical strategies) and problematic algorithmic bias (e.g., systematically degrading certain viewpoints' coherence). This distinction is subtle but crucial for responsible deployment.

10.3.5 Human-AI Collaborative Systems

Finally, future research should explore human-AI collaborative approaches to political discourse generation and analysis. Rather than fully automated generation, systems could provide politicians, speechwriters, and policy analysts with AI-assisted tools:

- **Rhetorical suggestion systems** that propose alternative phrasings consistent with party ideology
- **Consistency checking tools** that flag potential contradictions with previous statements
- **Counterargument generators** that help politicians anticipate opposing viewpoints
- **Framing analysis tools** that reveal how different phrasings activate different values

Such collaborative systems would leverage AI capabilities while maintaining human agency, judgment, and accountability. Research evaluating these systems' effectiveness and user acceptance would inform design of practical tools that enhance rather than replace human political deliberation.

10.4 Concluding Remarks

This thesis demonstrates that domain-specific fine-tuning of large language models enables generation of politically authentic discourse when evaluated through comprehensive, multi-dimensional frameworks. The success rate in achieving statistically significant improvements across diverse metrics confirms that fine-tuning on parliamentary speeches substantially enhances models' capacity to reproduce ideological nuances, party-specific rhetoric, and argumentative coherence characteristic of real political discourse.

The validation of novel political authenticity metrics (PSA, Party Align) establishes a foundation for quantitative assessment of ideological coherence in generated text, complementing traditional NLP measures that focus on linguistic quality alone. These metrics' demonstrated discriminative validity and fine-tuning responsiveness position them as valuable tools for both evaluation and analysis of political discourse.

Cross-context performance analysis revealed that architectural design, particularly context window length and attention mechanisms, shapes models' ability to capture extended argumentative structures and maintain ideological consistency. However, no single architecture dominated across all contexts; model-specific strengths and weaknesses suggest that ensemble approaches or task-specific model selection may optimize performance for diverse application scenarios.

The challenges identified in this research, in modeling technical topics, heterogeneous political groups, and ideological positions, illuminate both model limitations and characteristics of political language itself. These findings motivate continued investigation into how computational methods can capture the full complexity of political discourse while highlighting the importance of thoughtful evaluation frameworks that reflect domain-specific authenticity requirements.

This research contributes to that essential conversation by demonstrating what is technically feasible, establishing rigorous evaluation standards, and identifying critical considerations for responsible development and deployment of political discourse generation systems. The methods, metrics, and findings presented here provide a foundation for future work that must grapple with the profound implications of AI systems capable of mimicking human political expression.

Bibliography

- [1] 01. AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2025. doi:<https://doi.org/10.48550/arXiv.2403.04652>.
- [2] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023. doi:<https://doi.org/10.48550/arXiv.2305.13245>.
- [3] Goshi Aoki. Large language models in politics and democracy: A comprehensive survey, 2024. doi:<https://doi.org/10.48550/arXiv.2412.04498>.
- [4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. doi:<https://doi.org/10.48550/arXiv.2004.05150>.
- [5] Lorenzo Bocchi, Camilla Casula, and Alessio Palmero Aprosio. KEVLAR: The complete resource for EuroVoc classification of legal documents. In Felice Dell’Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 66–73, Pisa, Italy, December 2024. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL <https://aclanthology.org/2024.clicit-1.9/>.
- [6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15 (3), March 2024. ISSN 2157-6904. doi:[10.1145/3641289](https://doi.org/10.1145/3641289).
- [7] Mamata Das, Selvakumar K., and P. J. A. Alphonse. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset, 2023. doi:<https://doi.org/10.48550/arXiv.2308.04037>.
- [8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. doi:<https://doi.org/10.48550/arXiv.2305.14314>.
- [9] T. Erjavec, M. Kopp, N. Ljubešić, T. Kuzman, P. Rayson, P. Osenova, M. Ogrodniczuk, Ç. Çöltekin, D. Koržinek, K. Meden, J. Skubic, P. Rupnik, T. Agnoloni, J. Aires, S. Barkarson, R. Bartolini, N. Bel, M. Calzada Pérez, R. Dargis, S. Diwersy, M. Gavriilidou, R. van Heusden, M. Iruskieta, N. Kahusk, A. Kryvenko, N. Ligeti-Nagy, C. Magariños, M. Mölder, C. Navarretta, K. Simov, L. M. Tungland, J. Tuominen, J. Vidler, A. I. Vladu, T. Wissik, V. Yrjänäinen, and D. Fišer. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59(3):2071–2102, sep 2025. ISSN 1574-0218. doi:[10.1007/s10579-024-09798-w](https://doi.org/10.1007/s10579-024-09798-w).

- [10] Aaron Grattafiori and et al. The llama 3 herd of models, 2024. doi:<https://doi.org/10.48550/arXiv.2407.21783>.
- [11] Pema Gurung and Rupali Wagh. A study on topic identification using k means clustering algorithm: Big vs. small documents. *Advances in Computational Sciences and Technology*, 10(2): 221–233, 2017. ISSN 0973-6107. URL <http://www.ripublication.com>.
- [12] Erhard Hinrichs and Steven Krauwer. The CLARIN research infrastructure: Resources and tools for eHumanities scholars. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1525–1531, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.
- [13] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. doi:<https://doi.org/10.48550/arXiv.1904.09751>.
- [14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. doi:<https://doi.org/10.48550/arXiv.2106.09685>.
- [15] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023. ISSN 0020-0255. doi:[10.1016/j.ins.2022.11.139](https://doi.org/10.1016/j.ins.2022.11.139).
- [16] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005. ISSN 0001-4966. doi:[10.1121/1.2016299](https://doi.org/10.1121/1.2016299).
- [17] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023. doi:<https://doi.org/10.48550/arXiv.2310.06825>.
- [18] Valentin Kassarnig. Political speech generation, 2016. doi:<https://doi.org/10.48550/arXiv.1601.03313>.
- [19] Rahim Khan, Yurong Qian, and Sajid Naeem. Extractive based text summarization using kmeans and tf-idf. *International Journal of Information Engineering and Electronic Business*, 11:33–44, 05 2019. doi:[10.5815/ijeeb.2019.03.05](https://doi.org/10.5815/ijeeb.2019.03.05).
- [20] Marios Koniaris, Argyro Tsiipi, and Panayiotis Tsanakas. Parliabench: An evaluation and benchmarking framework for llm-generated parliamentary speech, 2025. URL <https://doi.org/10.48550/arXiv.2511.08247>.
- [21] Konstantinos Kyritsis, Charalampos M. Liapis, Nikolaos Spatiotis, Isidoros Perikos, and Michael Paraskevas. A comparative analysis of transformer models in zero-shot text classification. In *2024 15th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4, 2024. doi:[10.1109/IISA62523.2024.10786704](https://doi.org/10.1109/IISA62523.2024.10786704).

- [22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models, 2016. doi:<https://doi.org/10.48550/arXiv.1510.03055>.
- [23] Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, Alex Qian, Weixin Chen, Zhongkai Xue, Lichao Sun, Lifang He, Hanjie Chen, Kaize Ding, Zijian Du, Fangzhou Mu, Jiaxin Pei, Jieyu Zhao, Swabha Swayamdipta, Willie Neiswanger, Hua Wei, Xiyang Hu, Shixiang Zhu, Tianlong Chen, Yingzhou Lu, Yang Shi, Lianhui Qin, Tianfan Fu, Zhengzhong Tu, Yuzhe Yang, Jaemin Yoo, Jiaheng Zhang, Ryan Rossi, Liang Zhan, Liang Zhao, Emilio Ferrara, Yan Liu, Furong Huang, Xiangliang Zhang, Lawrence Rothenberg, Shuiwang Ji, Philip S. Yu, Yue Zhao, and Yushun Dong. Political-llm: Large language models in political science, 2024. doi:<https://doi.org/10.48550/arXiv.2412.06864>.
- [24] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities, 2024. doi:<https://doi.org/10.48550/arXiv.2408.13296>.
- [25] Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. Train once, test anywhere: Zero-shot learning for text classification, 2017. doi:<https://doi.org/10.48550/arXiv.1712.05972>.
- [26] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, jul 2024. doi:[10.48550/arXiv.2407.10671](https://doi.org/10.48550/arXiv.2407.10671).
- [27] Ludovic Rheault and Christopher Cochrane. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133, 2020. doi:[10.1017/pan.2019.26](https://doi.org/10.1017/pan.2019.26).
- [28] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. doi:<https://doi.org/10.48550/arXiv.2104.09864>.
- [29] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in LLM simulations of debates. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:[10.18653/v1/2024.emnlp-main.16](https://doi.org/10.18653/v1/2024.emnlp-main.16).
- [30] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshhev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus

Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. doi:<https://doi.org/10.48550/arXiv.2408.00118>.

- [31] Andrea Volkens, Judith Bara, Ian Budge, Michael McDonald, Robin Best, and Simon Franzmann. *Understanding and Validating the Left-Right Scale (RILE)*, pages 85–105. 11 2013. ISBN 9780199640041. doi:<https://doi.org/10.1093/acprof:oso/9780199640041.003.0006>.
- [32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi:[10.18653/v1/2020.emnlp-demos.6](https://doi.org/10.18653/v1/2020.emnlp-demos.6).
- [33] Wenhao Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashmi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madiha Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models, 2023. doi:<https://doi.org/10.48550/arXiv.2309.16039>.
- [34] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. doi:<https://doi.org/10.48550/arXiv.1904.09675>.
- [35] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen,

- Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. doi:<https://doi.org/10.48550/arXiv.2303.18223>.
- [36] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance, 2019. doi:<https://doi.org/10.48550/arXiv.1909.02622>.
- [37] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. doi:<https://doi.org/10.48550/arXiv.2306.05685>.
- [38] Wanzheng Zhu and Suma Bhat. Gruen for evaluating linguistic quality of generated text, 2020. doi:<https://doi.org/10.48550/arXiv.2010.02498>.
- [39] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models, 2018. doi:<https://doi.org/10.48550/arXiv.1802.01886>.