Validating Political Position Predictions of Arguments

Anonymous ACL submission

Abstract

Predicting political stance from argumentative discourse is a challenging NLP task, complicated by the subjective and continuous nature of political positioning. We evaluate 22 language models on this task using a dual-scale validation framework that combines pointwise and pairwise human annotation, addressing the limitations of pairwise evaluation alone for inherently subjective attributes. Using this framework, we construct and validate a large-scale knowledge base of political position predictions for 23,228 arguments from 30 *Question Time* debates. Pointwise evaluation shows moderate human–model agreement (Krippendorff's α=0.578), reflecting intrinsic subjectivity, while pairwise validation reveals substantially stronger alignment (α=0.86 for the best model), demonstrating that LLMs can reliably recover ordinal structure from subjective discourse. We release a validated argumentation knowledge base enabling graph-based reasoning and retrieval-augmented generation in political domains, and propose a generalisable evaluation methodology for subjective continuous NLP tasks where traditional categorical approaches fall short.

1 Introduction

Validating large-scale LLM predictions for subjective, continuous variables is hard. When the variable is political position – a continuous, contested quantity – the problem is compounded by a fundamental mismatch: pointwise human annotation scales well, but humans are cognitively ill-suited to precise pointwise judgements (Mosteller and Tukey, 1977; Likert, 1932) and absolute positioning on continuous scales (Thurstone, 1927; Tarlow et al., 2021). They are, however, reliable at comparative judgements (Kendall and Smith, 1940; Agresti, 2018). Full pairwise validation is prohibitive at scale (i.e., $O(n^2)$ comparisons versus $O(n)$ pointwise), so we propose a dual-scale framework that combines both: pointwise judgements to identify political arguments, and pairwise comparisons to validate their relative ordering.

We ground this approach in a knowledge base (KB) of 23,228 argumentative discourse units, or ADUs (Peldszus and Stede, 2013), extracted from 30 BBC *Question Time* debates, with political positions predicted by 22 LLMs and validated by over 1,500 crowdworkers. The resulting KB integrates argumentative structure with political position predictions, enabling fine-grained analyses and downstream applications, such as graph-based retrieval-augmented generation (RAG).

Argument mining (Lippi and Torroni, 2016b; Lawrence and Reed, 2019) and political science have largely evolved in isolation until recently (Lippi and Torroni, 2016a): existing work identifies *what* is being argued but ignores *where* those arguments fall on the political spectrum, resulting in no large-scale resource that jointly represents both. Our work fills this gap:

1. We release the first large-scale KB unifying locutions, ADUs, argumentative relations and political position predictions from 22 LLMs, enabling stance analysis at the granularity of individual arguments rather than whole texts or speakers.
2. We propose a novel dual-scale annotation methodology that combines pointwise judgement with comparative evaluation, and use it to validate political position predictions against comparative judgements from over 1,500 crowdworkers.
3. We demonstrate that pointwise LLM-predicted political positions can be meaningfully aligned with human comparative assessments, providing the first large-scale empirical grounding for this class of predictions.

2 Related Work

LLMs as Judges. The emergence of LLMs capable of nuanced language understanding and reasoning in unsupervised tasks has led to their adoption as evaluators of generated text. Recent work demonstrates that LLMs can reliably assess outputs produced by both humans and other models (Chen et al., 2023; Zhang et al., 2023; Chen et al., 2024b; Wang et al., 2024; Chen et al., 2024a), including instruction following, question answering, and domain-specific reasoning (Huang and Chang, 2023; Zhao et al., 2025). This shift reflects a broader trend in which generative models are used not only for evaluation but also for prediction and classification tasks that require interpreting subtle semantic and pragmatic cues.

Political Position Prediction with LLMs. Within political science and computational social science, LLMs have been used for political stance prediction at various granularities. Models were tasked with predicting politicians' political leanings using different ideological axes, such as gun control (Wu et al., 2023). GPT 3 and 4 were prompted to estimate the probability of a sentence being *conservative* or *liberal* in political party manifestos, taking the sentence-level mean as an analogue for the average position of each document (Ornstein et al., 2025). Multiple models were used to predict the political stance of sentences, using a scale from 0 (*extremely left*-wing) to 100 (*extremely right*-wing), within Tweets, manifestos, and policy speeches across several languages (Le Mens and Gallego, 2025).

Despite differences in scale and domain, this body of work predominately adopts pointwise model evaluation approaches, in which models assign absolute ideological scores or labels that are subsequently compared against aggregate human annotations. As we discuss below, such approaches may obscure systematic differences in judgement and place substantial cognitive demands on both human annotators and models.

Pairwise Comparison and Preference-Based Evaluation. A large body of psychological and decision-theoretic research has shown that pairwise comparison is cognitively simpler and more reliable than absolute rating for subjective judgement tasks (Thurstone, 1927). This insight underpins preference learning (Fürnkranz and Hüllermeier, 2003), which has become central to recent advances in LLM training. In particular, Reinforcement Learning from Human Feedback (RLHF) relies on pairwise human preferences to train reward models, with Bradley–Terry-style models commonly used to infer latent strength scores from comparison data (Christiano et al., 2017; Ouyang et al., 2022).

In contrast, existing work on LLM-based political position prediction has largely relied on pointwise scores rather than direct comparative judgements (Le Mens and Gallego, 2025). Recent studies have shown that aggregate correlation metrics can mask systematic disagreements between model outputs and human annotations, particularly when human uncertainty is high (Elangovan et al., 2025). Accordingly, chance-corrected agreement measures, such as Krippendorff's α, have been argued to provide a more appropriate basis for LLM–human comparison, as they explicitly account for agreement expected by chance (Haldar and Hockenmaier, 2025). These findings motivate the use of pairwise, preference-based evaluation for political stance judgements.

Political Argumentation Resources. Existing stance detection resources label texts categorically (*for*/*against* or discrete left–right) (Sim et al., 2013; Mohammad et al., 2016), while argument mining datasets focus on argumentative structure without modelling ideological position (Lippi and Torroni, 2016a; Menini et al., 2018; Haddadan et al., 2019; Visser et al., 2020; Mestre et al., 2021; Goffredo et al., 2022). Continuous stance measures combining ideological positioning with structured argumentative content remain unexplored.

3 Knowledge Base Construction

Figure 1 provides an overview of the pipeline we have developed for KB instantiation, prediction of political stances, and application via graph-based RAG. We explain each part of the pipeline below.

3.1 Data Sources and Preprocessing

We constructed a KB from 30 BBC Question Time debates[1] that were previously annotated for arguments by annotators with an inter-annotator agreement (Combined Argument Similarity Score) of 0.56 (Hautli-Janisz et al., 2022). The public corpora, represented in the Argument Interchange Format (AIF) (Chesñevar et al., 2006; Rahwan and

[1] https://corpora.aifdb.org/qt30 (Accessed on 4th December 2025)

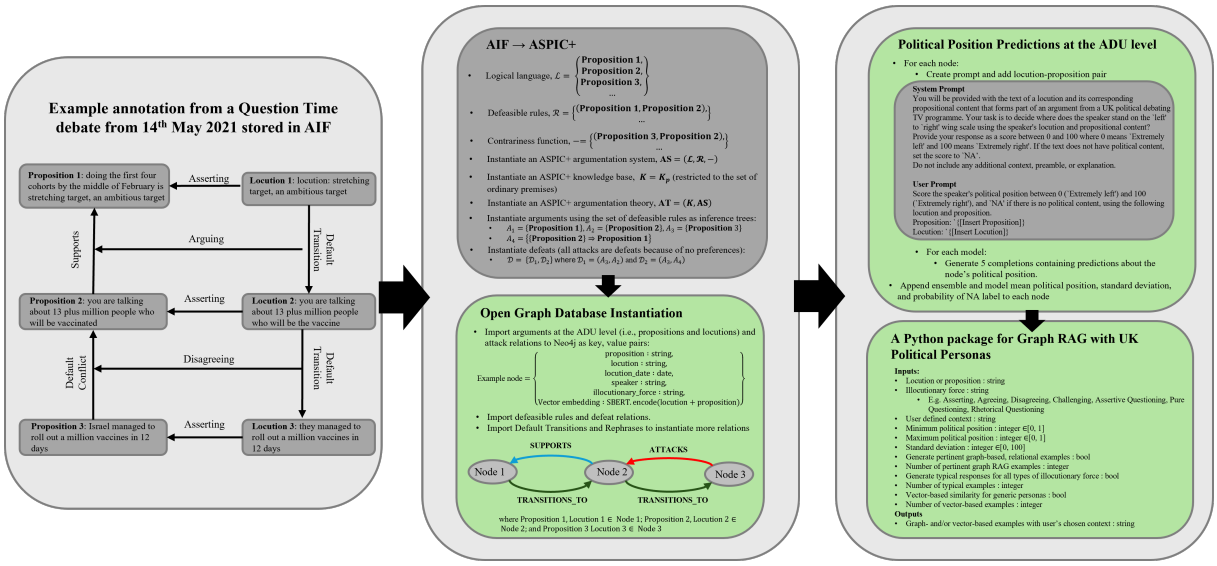

Figure 1: Overview of the methodology used to instantiate a structured argumentative KB containing political positions predictions for all arguments. Green boxes indicate components developed specifically for this study.

Reed, 2009), were converted into an ASPIC$^+$ argumentation theory following Bex et al. (2012), restricting the framework to ordinary premises and defeasible inference rules without preferences.

Arguments, defeat relations, and AIF dialogue transition structures were instantiated in a Neo4j (2012) graph database. We adopt the ASPIC$^+$ framework because it provides a formally rigorous bridge between natural-language argument structure and Dung-style abstract semantics (Prakken, 2010; Modgil and Prakken, 2014), supporting principled evaluation of argument acceptability. Granularity is important for graph-based RAG systems emulating political personas, which require ideologically aligned retrieval and coherent argumentative generation.

The resulting KB contains 23,228 arguments, consisting of locution-proposition pairs, and 50,905 relations comprised of support, attack, rephrase and chronological transition.

3.2 LLMs as Judges of Political Position

The 22 LLMs used to predict the political positions of arguments in the KB are visible in Table 5 (in App. A).

Building on the work of Le Mens and Gallego (2025), we extended sentence-level political stance predictions to both utterances and their corresponding ADU, interchangeably referred to as locution-proposition pairs or arguments from now on. Each model assigned a political stance score to all locution-proposition pairs on a 0–100 scale (left–right), or "NA" if an argument lacked political content. An example prompt is shown in Figure 1.

Each model scored every node in the graph five times to account for output variability even under near-deterministic settings (i.e., temperature set to 0, top_p equal to 0.1, and fixed seed for models that allowed parameter configuration). Prompting was executed using Golem (Blackwell, 2024) to ensure reproducible prompts with consistent parameter configurations at scale. We derived summary statistics for political position scores and the probability of an argument being labelled apolitical (or "NA") across repetitions, and stored these as properties of nodes in the KB.

3.3 Ensemble Construction and Aggregation

We constructed three ensembles of model predictions, each designed to isolate different aspects of model behaviour, prediction robustness and additional insights during evaluation. Summary statistics for political position score and probability of NA were calculated for each of the ensembles from the set of individual predictions from all models in those ensembles, and stored as properties for each node in addition to individual model predictions and summary statistics.

Ensemble 1 [E_1]: All ($n = 22$). A general-purpose baseline capturing average model behaviour across all models.

Ensemble 2 [E_2]: Reasoning Models ($n = 5$). Models with explicit reasoning or chain-of-thought

capabilities, to establish whether models designed to show their reasoning steps produce predictions more aligned with human judgements.

Ensemble 3 [E_3]: High-Confidence Models ($n = 12$). Models with high-confidence judgements, defined by the number of valid political stance predictions between 0–100 exceeding the number of NA predictions. Introduced after initial results indicated that smaller models were incapable of classifying arguments as NA, even for samples which were apolitical:[2]

$$E_3 = \{m \in E_1 \ : \ |\hat{y}_{pol}(m)| > |\hat{y}_{apol}(m)|\} \quad (1)$$

where m is a model, $\hat{y}_{pol}(m)$ is the number of arguments predicted as political by m, and $\hat{y}_{apol}(m)$ is the number arguments predicted NA by m.

4 Human Annotation and Validation Design

To validate model predictions of political positions introduced in Section 3, we made use of human crowdworkers to annotate arguments from the KB. Crowdworkers were recruited via Prolific.[3] All participants recruited resided in the UK, with English as first or primary language, and some form of higher education.

Following the structure of prompts presented to LLMs, we divided human annotation into two sequential tasks.

4.1 Pointwise Binary Classification of Political Sentiment

This task comprised pointwise binary classification of whether an argument was political or apolitical, validating model predictions while also providing a pool of high-confidence political arguments to validate the implicit ranking produced by model scores.

Annotation. We randomly sampled 1,000 arguments from our KB into three buckets, based on the mean probability of NA for Ensemble 3 ($\bar{\pi}^{(E_3)}$), in order to bias the dataset towards cases of high model confidence while still including some cases of low model confidence, as shown in Table 1.

All arguments obtained a majority label from three human annotations (binary labels), using a pool of 600 crowdworkers, each annotating five arguments. Two participants exhibiting invariance and atypical completion speed were flagged and replaced. We used *nominal* Krippendorff's α_n (Hayes and Krippendorff, 2007; Krippendorff, 2004) to measure inter-annotator agreement, accommodating partial annotator overlap and chance agreement.

Bucket	n	$\bar{\pi}^{(E_3)}$	Interpretation
H_{pol}	400	≤ 0.05	high confidence, political
L	200	$\in [0.45, 0.55]$	low confidence
H_{apol}	400	≥ 0.95	high confidence, apolitical

Table 1: Sampling buckets for pointwise binary classification of the presence of political sentiment.

Confidence-Based Dataset Partitioning. We partitioned $\mathcal{D}^{(\text{NA})}$ into two disjoint subsets: a high-confidence subset $\mathcal{D}^{(\text{NA})}_{\text{conf}} = \mathcal{D}^{(\text{NA})}_{\mathsf{H}_{\text{pol}}} \cup \mathcal{D}^{(\text{NA})}_{\mathsf{H}_{\text{apol}}}$, containing arguments models predicted as political or apolitical with high confidence, and an ambiguous subset $\mathcal{D}^{(\text{NA})}_{\text{ambig}} = \mathcal{D}^{(\text{NA})}_{\mathsf{L}}$, containing arguments with uncertain political status. By construction, $\mathcal{D}^{(\text{NA})}_{\text{conf}} \cap \mathcal{D}^{(\text{NA})}_{\text{ambig}} = \emptyset$ and $\mathcal{D}^{(\text{NA})}_{\text{conf}} \cup \mathcal{D}^{(\text{NA})}_{\text{ambig}} = \mathcal{D}^{(\text{NA})}$, enabling comparisons across confidence levels.

Evaluation. We evaluated presence of political sentiment using binary human majority labels and thresholded model predictions across $\mathcal{D}^{(\text{NA})}$, $\mathcal{D}^{(\text{NA})}_{\text{conf}}$, and $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$.

We calculated F1 score, precision, recall, and balanced accuracy (treating human labels as ground truth), in addition to α_n (framing parties as two equal-weight raters), allowing us to evaluate model performance relative to human reliability baselines.

4.2 Pairwise Comparison of Political Position

To validate the implicit ranking of arguments from pointwise model predictions, we presented human annotators with pairs of arguments, tasking them with annotating which one was more left- or right-leaning. Model-predicted political position scores were converted into pairwise comparisons and evaluated against human pairwise judgements via rankings derived from Bradley–Terry (BT) models (Bradley and Terry, 1952) and pairwise classification metrics across confidence levels.

Sampling Pairs. We sampled $n = 100$ arguments that were unanimously labelled as political by human annotators from $\mathcal{D}^{(\text{NA})}_{\mathsf{H}_{\text{pol}}}$, stratified according to position scores predicted by Ensemble

[2] E.g. some models assigned a score to nodes which contained a locution such as "I agree" which possessed no political sentiment.

[3] https://www.prolific.com (Accessed on 4th December 2025)

3. Continuous position predictions were discretised into deciles; bins corresponding to the ranges $0-10$ and $90-100$ were empty and therefore left excluded, yielding eight non-empty bins B. We denote the binned position score for argument i, repetition a, under model m as $s_{i,a}^{(m)} \in B = \{1, \ldots, 8\}$.

To do efficient comparison under resource constraints, we annotated a subset of 934 pairs drawn from the $\binom{100}{2} = 4,950$ possible argument pairings. Pair selection was guided by the mean predicted score $\bar{s}^{(E_3)}$ from Ensemble 3. Specifically, we sampled 44 intra-bin pairs from each bin (except for bin 8, where only 10 intra-bin pairs were available) and 22 inter-bin pairs. We denote the set of annotated pairs as $\mathcal{P} = \{(i_k, j_k) | k = 1, \ldots, 934\}$ where each element (i_k, j_k) corresponds to an ordered pair of arguments.

The resulting set of 934 pairs exceeds the $n \ln n \approx 460.5$ connectivity target (Negahban et al., 2012), where n=100, with balanced bin connections confirmed via Shannon (1948) entropy analysis (see App. B for details).

Win Matrices. We represent political position judgements as a hollow comparison matrix $W \in \mathbb{R}^{100\times100}$, where W_{ij} denotes the aggregated win count of argument i over j, with draws contributing 0.5 to both W_{ij} and W_{ji} (Davidson, 1970). For each model m, we constructed a dense win matrix $W^{(m)}$ from binned scores $s^{(m)}$ across all N[4] repetitions:

$$W_{ij}^{(m)} = \sum_{a,b=1}^{N} w\left(s_{i,a}^{(m)}, s_{j,b}^{(m)}\right) \quad w(x,y) = \begin{cases} 1 & x > y \\ 0.5 & x = y \\ 0 & x < y \end{cases} \tag{2}$$

For human annotations, 936 crowdworkers completed two symmetric tasks, judging which of two arguments was more left- or right-wing (or "equal"), with each pair annotated 3 times per task (6 total). Left-framed annotations were transposed and aggregated with right-framed responses, yielding three human win matrices: $W^{(H_L)}$, $W^{(H_R)}$, and $W^{(H)} = W^{(H_L),\top} + W^{(H_R)}$.

Pairwise Modelling. We fit BT models with the Iterative Luce Spectral Ranking (I-LSR) algorithm (Maystre and Grossglauser, 2015) provided in Python library `choix`,[5] to model pairwise judgements. This mapped comparisons onto a latent scale $\boldsymbol{\theta}$, where the probability of an outcome was given by $p(i > j) = \frac{e^{\theta_i}}{e^{\theta_i}+e^{\theta_j}}$. This process yielded a "smoothed" 100×100 probability matrix, imputing missing comparisons, and a ranking by sorting arguments in descending order of θ_i, where higher values indicate more right-wing positions. We use $\mathcal{R}^{(m)}$ to represent ranking determined by model m.

Confidence-based Dataset Partitioning. We partitioned $\mathcal{P}$ into four subsets $\mathcal{P}_{E_3,H}$ based on whether model ($W^{(E_3)}$) and human ($W^{(H)}$) win margins exceeded 0.25, yielding high- and low-confidence combinations for each: $\mathcal{P}_{1,1}$, $\mathcal{P}_{1,0}$, $\mathcal{P}_{0,1}$, and $\mathcal{P}_{0,0}$ (Table 2). Refer to Appendix C for details.

	E_3 Confident	E_3 Unconfident
H Confident	$\mathcal{P}_{1,1}$	$\mathcal{P}_{0,1}$
H Unconfident	$\mathcal{P}_{1,0}$	$\mathcal{P}_{0,0}$

Table 2: Subsets of $\mathcal{P}$ by model and human confidence.

Evaluation. We quantified agreement between rankings on our $n = 100$ arguments using Spearman's Footrule Distance (Diaconis and Graham, 1977) and Kendall's τ Distance (Kendall, 1938), min-max normalised to [0,1] and inverted to represent similarity metrics, which we represent as d_{footrule} and d_τ respectively. We also calculated ordinal Krippendorff's α_o to represent agreement. We calculate these metrics between each model ranking $\mathcal{R}^{(m)}$ and the human ranking $\mathcal{R}^{(H)}$, and use distance/agreement between aggregate human ranking $\mathcal{R}^{(H)}$ and rankings derived from the left- and right-framed annotation tasks ($\mathcal{R}^{(H_L)}$ and $\mathcal{R}^{(H_R)}$) to provide context to model results.

We calculated pairwise performance (macro-f1) of models, using human judgements as ground-truth, for all pairs receiving human annotation ($\mathcal{P}$). Labels represent a win or loss for the argument i in the pair (i, j) as judged by model m, and are calculated by thresholding $W_{ij}^{(m)}$ at 0.5. We present results across high- and low-confidence level subsets $\mathcal{P}_{E_3,H}$ for humans and models to breakdown how model performance differs across confidence levels.

To further contextualise model performance, we established two control baselines:

1. **Random Baseline:** no discriminative ability, i.e. assigns equal strength parameters (zero values) to all items, resulting in uniform probability matrices of 0.5.

[4] $N = 5$ for single LLMs; $N = 5 \times$ ensemble size for ensembles.

[5] https://choix.lum.li/ (Accessed on: 12th January 2026)

2. **Worst-Case Baseline:** systematically inverted predictions relative to human judgments, generated by negating the human aggregate model parameters and inverting probability matrices.

5 Experimental Results and Evaluation

5.1 Pointwise Annotation Study

Inter-Annotator Agreement. Overall, 48.5% of pointwise annotations were unanimous (3/3). To assess the reliability of the crowdsourced labels, we computed nominal Krippendorff's α_n. Across the full dataset $\mathcal{D}^{(\text{NA})}$ (n=1,000), inter-annotator agreement was low (α_n=0.305), indicating poor agreement amongst annotators.

When restricting the analysis to the confident subset $\mathcal{D}^{(\text{NA})}_{\text{conf}}$ (n=800), agreement increased slightly to α_n=0.317, whereas when only ambiguous cases were considered $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$ (n=200), agreement decreased to α_n=0.259. We note that α_n=0.483 for unanimously-labelled arguments (n=485) and α_n=0.436 for majority labels (n=515).

Inter-Model Agreement. Figure 2 reports α_n measuring agreement among model predictions under different dataset partitions. When models were evaluated exclusively on the ambiguous subset $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$, agreement is at or below chance (α_n=−0.048), indicating highly inconsistent predictions in regions of uncertainty.

When considering the full dataset $\mathcal{D}^{(\text{NA})}$, there was a substantial increase in agreement between models, such that models exhibited fair agreement (α_n=0.485), with the best model agreement observed in the confident partition $\mathcal{D}^{(\text{NA})}_{\text{conf}}$ (α_n=0.578), approaching the moderate agreement level.

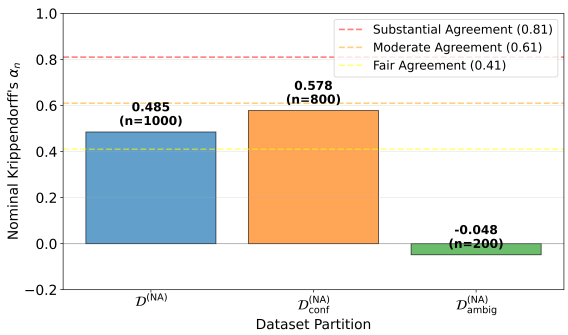

Figure 2: Nominal Krippendorff's α_n measuring inter-model agreement across data partitions.

Human–Model Agreement. Figure 3 shows the distribution of α_n between human annotations and model predictions across dataset partitions. Human–model agreement is highest on $\mathcal{D}^{(\text{NA})}_{\text{conf}}$, with a median of α_n=0.424. In contrast, agreement on $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$ is consistently negative, indicating systematic divergence between pointwise model predictions and human annotations.

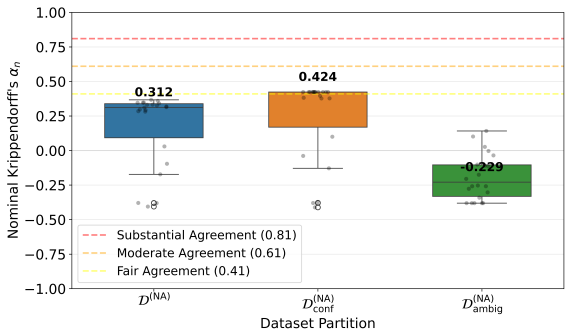

Figure 3: Distribution of human–model agreement across dataset partitions.

Model and Ensemble Performance. Figure 5 (in App. D.1) illustrates the relationship between human–model agreement and model performance (macro F1, micro F1 and balanced accuracy) across all models and ensembles, evaluated on $\mathcal{D}^{(\text{NA})}$, $\mathcal{D}^{(\text{NA})}_{\text{conf}}$, and $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$. Across all metrics, higher agreement with human judgements strongly predicts better performance. Macro F1 exhibits near-perfect correlations with agreement in both partitions and the full dataset, with similarly strong correlations observed for micro F1 and balanced accuracy.

Performance is consistently higher on $\mathcal{D}^{(\text{NA})}_{\text{conf}}$, demonstrating that models are most reliable on unambiguous instances. Moreover, $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$ exhibits substantial class imbalance as demonstrated by comparing macro and micro F1, likely contributing to degraded performance across all metrics and weaker agreement in this partition. These results establish human–model agreement as a robust proxy for model quality in pointwise political sentiment classification.

Beyond agreement-based analysis, we explicitly evaluated predictive performance for the three ensembles introduced in Section 3.3, alongside the best performing models (Table 3). All three ensembles demonstrate consistent performance, with near-identical metrics across most conditions. Again, dataset partitioning had a substantial impact on performance across all ensembles and models,

with results from $\mathcal{D}^{(\text{NA})}_{\text{conf}}$ consistently outperforming both ambiguous cases and the full dataset.

Model	Macro F1			Micro F1			Bal. Acc.		
	$\mathcal{D}^{(\text{NA})}$	$\mathcal{D}^{(\text{NA})}_{\text{conf}}$	$\mathcal{D}^{(\text{NA})}_{\text{ambig}}$	$\mathcal{D}^{(\text{NA})}$	$\mathcal{D}^{(\text{NA})}_{\text{conf}}$	$\mathcal{D}^{(\text{NA})}_{\text{ambig}}$	$\mathcal{D}^{(\text{NA})}$	$\mathcal{D}^{(\text{NA})}_{\text{conf}}$	$\mathcal{D}^{(\text{NA})}_{\text{ambig}}$
Ensemble 1	0.660	0.712	0.308	0.660	0.714	0.445	0.680	0.720	0.500
Ensemble 2	0.660	0.712	0.348	0.660	0.714	0.445	0.678	0.720	0.493
Ensemble 3	0.675	0.712	0.528	0.677	0.714	0.530	0.681	0.720	0.530
GPT 4o	0.683	0.712	0.570	0.685	0.714	0.570	0.691	0.720	0.575
DeepSeek-V3	0.680	0.712	0.550	0.681	0.714	0.550	0.688	0.720	0.559
Grok 2	0.673	0.712	0.512	0.674	0.714	0.515	0.682	0.720	0.530
Gemini 2.5 Flash	0.671	0.712	0.481	0.671	0.714	0.500	0.683	0.720	0.527
GPT 3.5 Turbo	0.655	0.691	0.497	0.655	0.691	0.510	0.672	0.705	0.534
GPT 4.5	0.673	0.712	0.448	0.681	0.714	0.550	0.674	0.720	0.509
Llama 4 Maverick	0.666	0.712	0.441	0.666	0.714	0.475	0.680	0.720	0.508
Gemini 1.5 Pro	0.660	0.712	0.444	0.663	0.714	0.460	0.663	0.720	0.447
Qwen 3	0.663	0.712	0.435	0.663	0.714	0.460	0.676	0.720	0.489
Claude 3.7 Sonnet	0.666	0.712	0.411	0.673	0.714	0.510	0.666	0.720	0.471

Table 3: Ensemble and top ten performing models across $\mathcal{D}^{(\text{NA})}$, as well as $\mathcal{D}^{(\text{NA})}_{\text{conf}}$ and $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$ partitions.

Discussion. Across inter-annotator, inter-model, and human–model analyses, a consistent pattern emerges: agreement is systematically higher on $\mathcal{D}^{(\text{NA})}_{\text{conf}}$ and degrades sharply when ambiguous instances are included. Human annotators exhibit low overall agreement, reflecting the inherent subjectivity of pointwise political sentiment annotation, while models achieve substantially higher consistency on the same confident subset.

Both humans and models struggle on $\mathcal{D}^{(\text{NA})}_{\text{ambig}}$, indicating a region of genuine semantic uncertainty rather than stochastic prediction noise. The convergence between human disagreement and model uncertainty provides empirical evidence that human–model agreement is a reliable indicator of model prediction quality. Excluding ambiguous instances yields marked improvements in both agreement and downstream performance, reinforcing the conclusion that pointwise model predictions can approximate human judgements of pointwise binary classification of political sentiment.

5.2 Validating Model Predictions using Pairwise Human Annotations

Inter-Annotator Agreement. We measured ordinal Krippendorff's α_o between the inverted left-framed and right-framed annotation tasks to assess the reliability of the crowdsourced pairwise comparisons. From a carefully selected subset of 100 political arguments, we constructed 934 samples and obtained α_o=0.889, indicating substantial agreement between framing and supporting the robustness of our pairwise annotation protocol.

Most pairs received two distinct labels (57-63%), with complete agreement in 23–28% of cases and full disagreement rare (<15%). Refer to Appendix E.1 for more discussion on this.

Model and Ensemble Performance, Agreement and Distance Metrics. To evaluate alignment between model-inferred political rankings and human judgements under varying levels of uncertainty, we report agreement and performance metrics over the full dataset $\mathcal{P}$ (Fig. 6 in App. E.2) and over four conditional subsets (Fig. 7 in App. E.2): $\mathcal{P}_{1,1}$ (high-confidence model predictions and human annotations), $\mathcal{P}_{1,0}$ (high-confidence model predictions, low-confidence human annotations), $\mathcal{P}_{0,1}$ (low-confidence model predictions, high-confidence human annotations), and $\mathcal{P}_{0,0}$ (low-confidence predictions on both sides). We report Spearman's d_{footrule}, Kendall's τ, ordinal α_o, and macro F1 for all ensembles and the top-performing individual models (Table 4).

Across all models, ranking agreement with humans under $\mathcal{P}$ remains substantially below human inter-annotator agreement. Individual models achieve α_o values that ranged from 0.52 to 0.80, corresponding to moderate agreement but falling short of the substantial agreement threshold (α_o=0.81; Fig. 6 in App. E.2). This gap persists even among the strongest propriety models, indicating that inferring political orderings from pointwise judgements remains a challenging task.

Conditioning on high-confidence predictions markedly improves performance. Under $\mathcal{P}_{1,1}$, several models – including GPT 4.5, o3 Mini, DeepSeek-R1 and Claude 3.5 Haiku – approached or exceeded $\alpha_o \approx 0.85$, placing them close to the boundary of agreement between humans across the left- and right-framed tasks (Table 4). As shown in Figure 7 (in App. E.2), agreement increased and variance decreased in subsets where human confidence is high, whereas low-confidence subsets exhibited weaker and more variable alignment.

Human Upper Bound. Human annotations established a clear upper bound on achievable performance. Aggregate human inter-annotator agreement reached α_o=1.0 by construction, while individual left- and right-framed annotations retained extremely high agreement with the aggregate ranking ($\alpha_o \geq 0.966$ for $\mathcal{P}$; Table 4). Notably, conditioning on high-confidence judgements ($\mathcal{P}_{1,1}$) decreased the agreement slightly but increased annotation performance, yielding perfect macro F1 for both annotation conditions.

Ensembles versus Individual Models. Under the full distribution $\mathcal{P}$, ensemble methods consistently outperformed individual models. All three

Model	d_{footrule}		d_τ		α_o		Macro F1	
	$\mathcal{P}$	$\mathcal{P}_{1,1}$	$\mathcal{P}$	$\mathcal{P}_{1,1}$	$\mathcal{P}$	$\mathcal{P}_{1,1}$	$\mathcal{P}$	$\mathcal{P}_{1,1}$
Human (agg)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Human (left)	0.904	0.867	0.934	0.907	0.974	0.956	0.827	1.000
Human (right)	0.886	0.873	0.922	0.914	0.966	0.960	0.790	1.000
Ensemble 1	0.726	0.721	0.810	0.800	0.808	0.789	0.603	0.492
Ensemble 2	0.721	0.682	0.804	0.782	0.803	0.760	0.622	0.617
Ensemble 3	0.728	0.657	0.806	0.756	0.799	0.692	0.583	0.492
o3 Mini	0.684	0.779	0.779	0.837	0.766	0.852	0.644	**0.723**
GPT 4.5	0.696	**0.806**	**0.793**	**0.852**	**0.783**	0.849	0.630	0.492
DeepSeek-R1	0.691	0.774	0.781	0.840	0.753	0.849	0.595	0.617
Qwen QwQ	0.661	0.760	0.770	0.819	0.747	0.813	0.621	0.673
GPT 4o Mini	0.661	0.780	0.767	0.829	0.730	0.827	**0.645**	0.570
Claude 3.5 Haiku	0.679	0.785	0.772	0.842	0.734	**0.858**	0.640	0.483
DeepSeek-V3	**0.698**	0.753	0.777	0.819	0.739	0.827	0.595	0.484
Gemini 1.5 Pro	0.612	0.743	0.734	0.806	0.655	0.778	0.610	0.585
Random Baseline	0.333	0.333	0.500	0.500	0.000	0.000	0.500	0.500
Worst-case Baseline	0.000	0.000	0.000	0.000	-1.000	-1.000	0.000	0.000

Table 4: Model performance metrics, by metric and dataset. Features all models that acheived top-5 performance for each metric. Bold cells indicate highest performing LLM for given column.

ensembles achieved $\alpha_o \approx 0.80$, outperforming most individual models and narrowing (but not closing) the human–model gap (Fig. 6 in App. E.2 and Table 4). Interestingly, this advantage diminishes under $\mathcal{P}_{1,1}$. When attention is restricted to high-confidence predictions, several individual models match or exceed ensemble performance on ranking agreement and macro F1.

Ranking Agreement v. Pointwise Model Accuracy. A key insight from from Table 4 (Fig. 6 in App. E.2) is the partial decoupling of ranking agreement and macro F1. Some models (e.g., GPT 4o Mini) achieved a strong macro F1 under $\mathcal{P}$ despite lower ranking agreement, while others prioritise ordinal consistency at the expense of pointwise accuracy. Conditioning on $\mathcal{P}_{1,1}$ sharpened this trade-off: macro F1 generally increased for confident subsets but gains in α_o were more pronounced, highlighting that ordinal ranking benefits more from uncertainty filtering than categorical accuracy does.

Discussion. Our pairwise annotation study clarifies how pointwise model-generated political judgements relate to human comparative reasoning under uncertainty. Rather than framing validation solely as agreement maximisation, our results show that confidence and aggregation play a central role in shaping the recoverable ordinal structure.

The gap between human and model agreement under the full distribution is largely attributable to uncertainty. While models exhibit only moderate alignment overall, conditioning on high-confidence judgements reveals substantial agreement with the human aggregate ranking for several models. This pattern suggests that disagreement is driven less by systematic ideological bias than by structurally ambiguous regions of the ordering that also challenge human annotators. Accordingly, model outputs encode a partially correct ordinal structure, with uncertainty concentrated in weakly constrained areas of the comparison graph.

Our findings further distinguish ordinal consistency from pointwise accuracy. Ordinal agreement and macro F1 are only weakly coupled, and improvements in one do not necessarily imply gains in the other. Confidence filtering disproportionately benefits ordinal structure, reinforcing the view that global rankings and local classifications capture distinct epistemic aspects of the task.

Aggregation effects follow a similar pattern. Ensembles improve ordinal agreement under the full distribution by stabilising rankings in ambiguous regions, but offer limited advantage once uncertainty is reduced, where strong individual models perform comparably. Finally, despite being elicited without explicit comparisons, pointwise model predictions recover a substantial portion of the ordinal information expressed in human pairwise judgements, particularly in high-confidence cases. This supports their use as a scalable proxy for human comparative annotation when exhaustive pairwise judgements are impractical.

6 Conclusions and Future Work

Disagreement in pointwise political position annotation reflects the inherent difficulty of absolute judgements, not model unreliability. Pairwise comparison yields substantially higher agreement and confirms that pointwise LLM predictions recover meaningful ordinal political structure, especially under high-confidence conditions. The validated subset of the KB thus help to delineate where political positioning can be interpreted reliably and where uncertainty is intrinsic.

We release a large-scale KB that integrates attack and support relations, dialogue structure, political position predictions from 22 LLMs, ensemble aggregates, and uncertainty estimates at the ADU-level. The KB was validated by over 1,500 crowdworkers and should support a wealth of future work, such as analyses of how ideology relates to attack and support relations in dialogical argumentation.

Future work will extract political positions from ADUs in new domains, e.g. international political systems that do not have a dichotomous left-right division, and generate politically aligned personas using the shared resource.

Limitations

We note some limitations. 1) Pointwise political position annotations is inherently difficult for humans, as seen in substantially lower human–human agreement compared to the pairwise study, reinforcing prior findings that absolute judgments on continuous ideological scales are cognitively demanding. This limits the reliability of pointwise labels as a gold standard and motivates our dual-scale validation design, but does not eliminate the underlying subjectivity of the task. 2) Ensemble construction (esp. Ensemble 3) relies on estimating the probability that an argument is apolitical, requiring multiple model invocations per sample. At runtime, it is not known whether an argument will be unambiguous, making the approach more computationally costly and restricting it in low-latency settings. Relatedly, ensembles primarily improve performance in ambiguous regions of the comparison space; confident cases are often handled comparably well by individual models, though ambiguity itself cannot be identified prior to inference. 3) Our pairwise validation relies on discretisation of continuous model outputs and on BT regularisation choices, both of which may obscure finer-grained ordinal differences. 4) Pairwise rankings are derived from a relatively small subset of arguments (100) drawn from a single political discourse domain, which may limit generalisability to other corpora or ideological axes.

Ethical Considerations

This study received ethical approval from the University of X (Ref: #####), University of Y (Ref: #####) and Funder Z (Ref: #####). Human crowdworkers were recruited through the online research platform Prolific. All participants provided informed consent prior to undertaking the annotation tasks, in accordance with institutional guidelines. No personally identifiable information was collected by the research team, and participant anonymisation was managed by the Prolific platform. Participants were compensated at an average rate of £9.76 per hour across both left- and right-framed annotation tasks. The annotation guidelines used pointwise binary annotation study, as well as the left- and right-framed tasks, can be found in Appendix F.

Data and Code Availability

We release the structured KB derived from 30 episodes of BBC Question Time, including arguments, their associated political position scores and relations, together with the associated Prolific-based annotation dataset and anonymised demographic metadata. The resource is provided as a database dump and Dockerfile for reproducible deployment, with supporting code available on our GitHub repository: `https://github.com/anonymous-argumentation/Validating-Political-Position-Predictions-of-Arguments`. All data and code are released under the MIT License.

Acknowledgements

The work reported in this paper was supported by funding from Funder Z. We also acknowledge support from Microsoft Research's Accelerating Foundation Models Research programme, which provided Azure resource to run a selection of the LLMs used in the experiments that are reported in this paper.

References

Alan Agresti. 2018. Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 41(2):287–297.

Anthropic. 2024. Claude 3.5 Haiku. Model identifier: claude-3-5-haiku-20241022.

Anthropic. 2025. Claude 3.7 Sonnet. Model identifier: claude-3-7-sonnet-20250219.

F. Bex, S. Modgil, H. Prakken, and C. Reed. 2012. On logical specifications of the Argument Interchange Format. *Journal of Logic and Computation*, 23(5):951–989.

Robert Blackwell. 2024. Golem – robblackwell/golem: v0.0.1-alpha, doi.org/10.5281/zenodo.14035711.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024b.

HuatuoGPT-II, one-stage training for medical adaption of LLMs. *https://arxiv.org/abs/2311.09774*.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Phoenix: Democratizing ChatGPT across languages. *https://arxiv.org/abs/2304.10453*.

Carlos Chesñevar, Jarred Mcginnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *Knowledge Eng. Review*, 21:293–316.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Roger R. Davidson. 1970. On extending the Bradley-Terry Model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.

DeepSeek-AI. 2025a. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

DeepSeek-AI. 2025b. DeepSeek-V3 Technical Report. *https://arxiv.org/abs/2412.19437*.

Persi Diaconis and R. L. Graham. 1977. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Johannes Fürnkranz and Eyke Hüllermeier. 2003. Pairwise preference learning and ranking. In *Machine Learning: ECML 2003*, pages 145–156, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *https://arxiv.org/abs/2403.05530*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *https://arxiv.org/abs/2507.06261*.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Rajarshi Haldar and Julia Hockenmaier. 2025. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24986–25004, Suzhou, China. Association for Computational Linguistics.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3291–3300. European Language Resources Association.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

M. G. Kendall and B. Babington Smith. 1940. On the method of paired comparisons. *Biometrika*, 31(3-4):324–345.

K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Business & Economics. Sage.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Gaël Le Mens and Aina Gallego. 2025. Positioning political texts with large language models by asking and averaging. *Political Analysis*, page 1–9.

R. Likert. 1932. *A Technique for the Measurement of Attitudes*. Number nos. 136-165 in A Technique for the Measurement of Attitudes. Columbia university.

Marco Lippi and Paolo Torroni. 2016a. Argument mining from speech: Detecting claims in political debates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Marco Lippi and Paolo Torroni. 2016b. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2).

Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of plackett–luce models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meta AI. 2024a. The llama 3 herd of models. *https://arxiv.org/abs/2407.21783*.

Meta AI. 2024b. Llama 3.2. *https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/*. Large language model accessed locally; model snapshot llama3.2:3b.

Meta AI. 2024c. Llama 3.3. *https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/*. Large language model accessed via Azure's API; model snapshot llama3.3:70b.

Meta AI. 2025. Llama 4 maverick. *https://ai.meta.com/blog/llama-4-multimodal-intelligence/*. Large language model accessed via Azure's API; model snapshot llama-4-maverick.

Microsoft. 2024. Phi-4 technical report. *https://arxiv.org/abs/2412.08905*. Large language model accessed via OpenRouter's API; model snapshot microsoft/phi-4.

Mistral AI. 2023. Mistral 7b. *https://arxiv.org/abs/2310.06825*. Large language model accessed locally; model snapshot mistral:7b.

Sanjay Modgil and Henry Prakken. 2014. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

F. Mosteller and J.W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2012. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Neo4j. 2012. Neo4j - the world's leading graph database. *http://neo4j.org/*.

OpenAI. 2024a. GPT-3.5 Turbo. Large language model accessed via Azure's API to OpenAI; model snapshot gpt-3.5-turbo-0125.

OpenAI. 2024b. GPT-4 Turbo. Large language model accessed via Azure's API to OpenAI; model snapshot gpt-4-turbo-2024-04-09.

OpenAI. 2024c. GPT-4o. Large language model accessed via Azure's API to OpenAI; model snapshot gpt-4o-2024-08-06.

OpenAI. 2024d. GPT-4o Mini. Large language model accessed via Azure's API to OpenAI; model snapshot gpt-4o-mini-2024-07-18.

OpenAI. 2025a. GPT-4.5 Preview. Preview large language model accessed via Azure's API to OpenAI; model snapshot gpt-4.5-preview-2025-02-27.

OpenAI. 2025b. o3-mini. *https://openai.com/index/openai-o3-mini/*. Large language model accessed via Azure's API to OpenAI; model snapshot o3-mini-2025-01-31.

Joseph T. Ornstein, Elise N. Blasingame, and Jake S. Truscott. 2025. How to train your stochastic parrot: large language models for political texts. *Political Science Research and Methods*, 13(2):264–281.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Henry Prakken. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124.

Qwen Team. 2024. Qwq. *https://qwen.ai/blog?id=qwq-32b-preview*. Large language model accessed via OpenRouter's API; model snapshot qwq-32b.

Qwen Team. 2025. Qwen3 technical report. *https://arxiv.org/abs/2505.09388*. Large language model accessed via OpenRouter's API; model snapshot qwen3-235b-a22b.

Iyad Rahwan and Chris Reed. 2009. The argument interchange format. In *Argumentation in Artificial Intelligence*, pages 383–402.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA. Association for Computational Linguistics.

Kevin R. Tarlow, Daniel F. Brossart, Alexandra M. McCammon, Alexander J. Giovanetti, M. Camille Belle, and Joshua Philip. 2021. Reliable visual analysis of single-case data: A comparison of rating, ranking, and pairwise methods. *Cogent Psychology*, 8(1):1911076.

L. L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. Large language models can be used to estimate the latent positions of politicians. *https://arxiv.org/abs/2303.12057*.

xAI. 2024. Grok 2. *https://x.ai/news/grok-1212*. Large language model accessed via Azure's API to xAI; model snapshot grok-2-1212.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. *https://arxiv.org/abs/2305.15075*.

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, and 3 others. 2025. A survey of large language models. *https://arxiv.org/abs/2303.18223*.

A Models and Ensembles

The models used to predict political positions of all the arguments in the KB, and the ensembles created afterwards, are presented in Table 5.

Model	Official Name	E_1	E_2	E_3
Claude 3.5 Haiku	`claude-3-5-haiku-20241022` (Anthropic, 2024)	✓		
Claude 3.7 Sonnet	`claude-3-7-sonnet-20250219` (Anthropic, 2025)	✓	✓	✓
DeepSeek-R1	`deepseek-r1-0528` (DeepSeek-AI, 2025a)	✓	✓	✓
DeepSeek-V3	`deepseek-v3-0324` (DeepSeek-AI, 2025b)	✓		✓
Gemini 1.5 Pro	`gemini-1.5-pro-002` (Gemini Team, 2024)	✓		✓
Gemini 2.5 Flash	`gemini-2.5-flash-preview` (Gemini Team, 2025)	✓		✓
GPT 3.5 Turbo	`gpt-3.5-turbo-0125` (OpenAI, 2024a)	✓		
GPT 4 Turbo	`gpt-4-turbo-2024-04-09` (OpenAI, 2024b)	✓		✓
GPT 4o	`gpt-4o-2024-08-06` (OpenAI, 2024c)	✓		✓
GPT 4o Mini	`gpt-4o-mini-2024-07-18` (OpenAI, 2024d)	✓		
GPT 4.5	`gpt-4.5-preview-2025-02-27` (OpenAI, 2025a)	✓		✓
Grok 2	`grok-2-1212` (xAI, 2024)	✓		✓
Llama 3.1:8b	`llama3.1:8b` (Meta AI, 2024a)	✓		
Llama 3.1:405b	`llama-3.1-405b-instruct` (Meta AI, 2024a)	✓		
Llama 3.2:3b	`llama3.2:3b` (Meta AI, 2024b)	✓		
Llama 3.3:70b	`llama-3.3-70b-instruct` (Meta AI, 2024c)	✓		
Llama 4 Maverick	`llama-4-maverick` (Meta AI, 2025)	✓		✓
Mistral:7b	`mistral:7b` (Mistral AI, 2023)	✓		
o3 Mini	`o3-mini-2025-01-31` (OpenAI, 2025b)	✓	✓	
Phi 4	`microsoft/phi-4` (Microsoft, 2024)	✓		✓
Qwen 3	`qwen3-235b-a22b` (Qwen Team, 2025)	✓	✓	✓
Qwen QwQ	`qwq-32b` (Qwen Team, 2024)	✓	✓	
	Total	22	5	12

Table 5: Models included in each ensemble.

B Pairwise Sample Connectivity and Entropy

The set of 934 pairwise samples exceeds the $n \ln n \approx 460.5$ target for pairwise connections (Negahban et al., 2012), where n=100 for this study. We verified full connectivity of the resulting graph, and confirmed even distributions of connections by computing node connection entropy values, defined as the Shannon (1948) entropy $H(v_i)$ of an item v_i (Figure 4). This is calculated as the sum of proportions of comparisons f between v_i and each bin $b \in B$ ($f_{i,b}$) weighted by their logarithms, shown in Equation (3). With an ideal upper bound $\log_2(8) \approx 3$ (uniform distribution across all eight bins), the median value of the resulting distribution was 2.5, with more than 60% of scores falling within 2.2–2.8, indicating balanced connections (lower values relate to bins with availability constraints).

$$H(v_i) = -\sum_{b \in B} f_{i,b} \log_2(f_{i,b}) \quad (3)$$

C Pairwise Confidence-based Dataset Partitioning

We assigned pairs from $\mathcal{P}$ into subsets based on the confidence of human and model (using ensemble 3) judgements, to account for uncertainty introduced

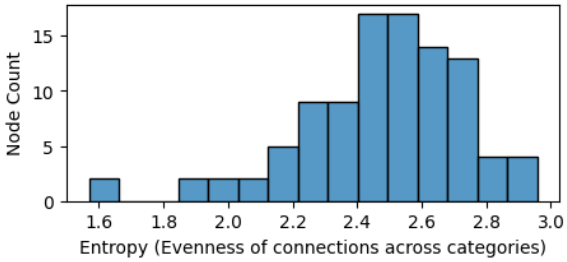

Figure 4: Distribution of Shannon (1948) entropy across the set of 934 pairwise comparisons.

across model repetitions and through the 6 multi-class human annotations per pair.

A given pair i, j in $\mathcal{P}$ was assigned two confidence values ($|W_{ij} - 0.5| \geq 0.25 \in \{0, 1\}$), for $W^{(E_3)}$ and $W^{(H)}$, with 1 indicating high-confidence. This provided four subsets $\mathcal{P}_{E_3,H} \subset \mathcal{P}$ across different combinations of model and human confidence (Table 2).

D Pointwise Annotation Study Results

D.1 Model and Ensemble Performance

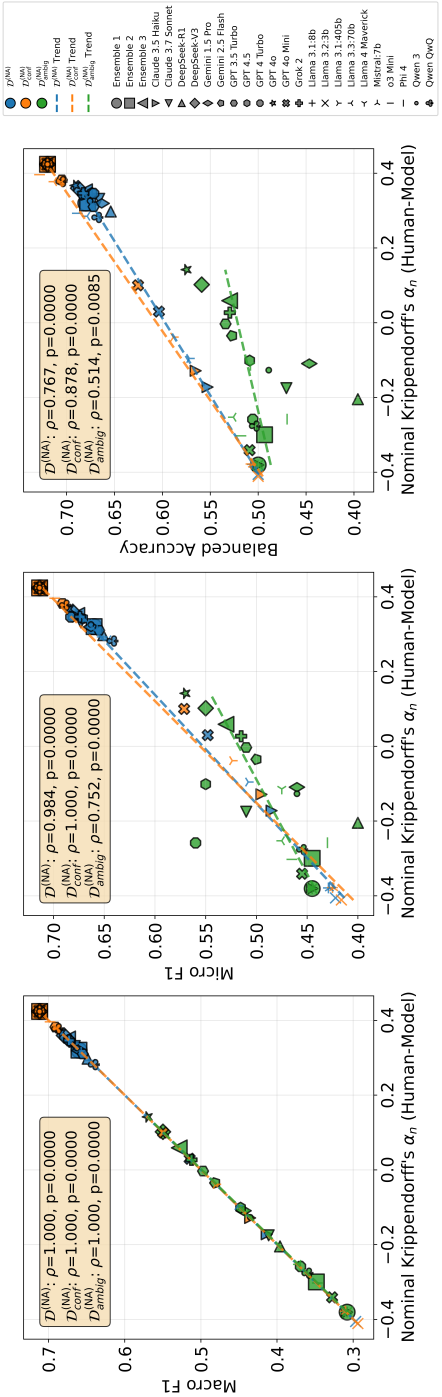

Figure 5: Relationship between human–model agreement and model performance metrics.

E Pairwise Annotation Study Results

E.1 Distribution of Human Agreement in the Pairwise Annotation Study

Table 6 reports the distribution of the number of unique labels assigned per pair across annotators for the left-framed and right-framed pairwise annotations tasks. In both conditions, the majority of pairs received two distinct labels (57.3% left-framed; 62.8% right-framed), indicating broader annotator consensus with limited disagreement. Pairs exhibiting complete agreement – characterised by a single unique label – account for a substantial minority of comparisons (27.9% and 22.5%, respectively), while complete disagreement (three unique labels) is rare and occurred at comparable rates across framings (14.8% left-framed; 14.7% right-framed). The close correspondence between distributions further confirms that inter-rater agreement patterns were stable across annotation framings.

	Left-framed		Right-framed	
Number of unique labels	%	n	%	n
1	27.9	261	22.5	210
2	57.3	535	62.8	587
3	14.8	138	14.7	127

Table 6: Unique labels per pair across annotation tasks.

E.2 Model and Ensemble Performance versus Ordinal Agreement

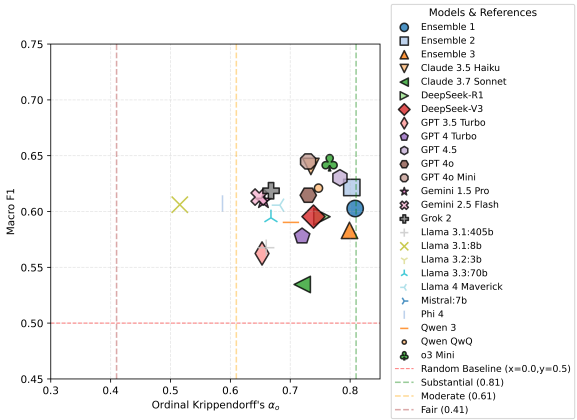

Figure 6: Macro F1 plotted against ordinal human–model agreement for all models and ensembles under $\mathcal{P}$.

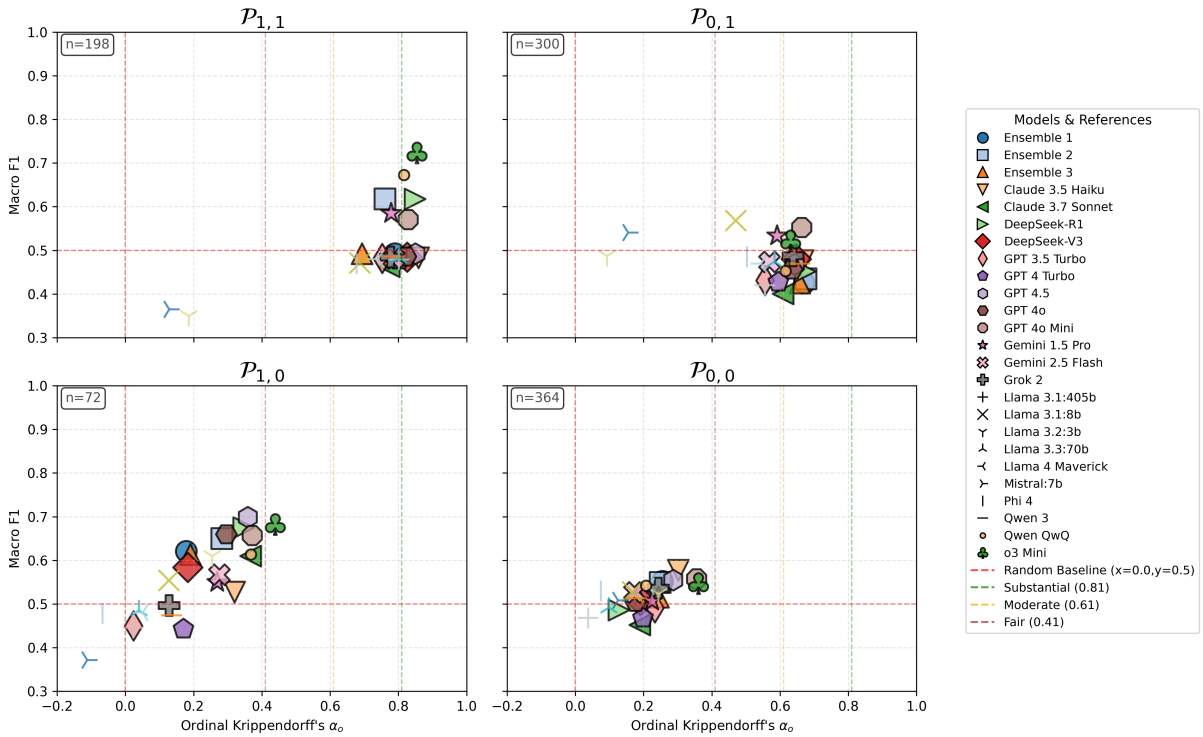

Figure 7: Macro F1 plotted against ordinal human–model agreement for all models and ensembles under four distributions: $\mathcal{P}_{1,1}$, which conditioned on high-confidence model predictions and human annotations; $\mathcal{P}_{1,0}$, which conditioned on high-confidence model predictions and low-confidence human judgements; $\mathcal{P}_{0,1}$, which conditioned on low-confidence model predictions and high-confidence human annotations; and $\mathcal{P}_{0,0}$, which conditioned on low-confidence model predictions and human annotations.

F Annotation Guidelines

F.1 Pointwise Annotation Study

Task Introduction You will be looking at statements derived from comments made in political debates. The two statements (labelled *Proposition* and *Locution*) relate to the same spoken comment: *Locution* is a direct transcription of the comment, while *Proposition* is a rephrased version of the same comment to include relevant context and names.

Taking these two pieces of text together, you are required to determine whether the spoken comment has a political leaning or not.

Task Steps

1. Read both the *Proposition* and the *Locution*.
2. Decide whether the comment has a political leaning or not, answering *Yes* or *No*.
   - **Yes**: the comment takes a particular stance on a specific policy area or the actions of government or politicians.
   - **No**: the comment may (or may not) discuss a specific policy area or government or politicians, but ultimately presents no political stance.

Examples

"Yes" examples:

- "Most governments have been too slow to act on the border issue."
- "We need to make sure we exploit the potential of wind farms and offshore as well."
- "Where there are breaches on the rules, nothing seems to happen in this government."

"No" examples:

- "It is incredibly important that children are able to learn."
- "We need to establish a way and a means to go forward with this."
- "China, India, Russia are participating in the 100 million financing for emerging markets."

F.2 Left- and Right-Framed Annotation Studies

Task Introduction You will be looking at statements derived from comments made in political debates. The two statements are presented as they were originally spoken and in *propositional* form, i.e. a rephrased version of the same comment to include relevant context and names.

Taking these two pieces of text together, you are required to compare Statement 1 and Statement 2, and determine which has a more <LEFT or RIGHT/>-wing political leaning (or whether they are more or less equal).

Task Steps

1. Read Statement 1 and Statement 2.
2. Decide which has a more <LEFT or RIGHT/>-wing political leaning, selecting the corresponding option, or *Equal* if neither is clearly more <LEFT or RIGHT/>-wing than the other.

Examples

Left-wing examples:

- "We should have spent the money making sure kids and the vulnerable in society had more."
- "We need to understand the transatlantic slave trade and its impact on people of African descent today."
- "It's a never ending Tory elitist rule, and that's how it's perceived in Scotland."

Right-wing examples:

- "Restrictions will destroy businesses and confidence across the country."
- "I think people will get tired of taking the knee, much like they are tired of woke."
- "The Labour party is unable to put a stance on what it will do to protect communities here from the terrorists."