An Argumentation Framework for Graph RAG with UK Political Personas

Anonymous ACL submission

Abstract

We present an argumentation framework that was instantiated using argumentative data from 30 debates that aired on the BBC television politics programme *Question Time* throughout 2020 and 2021. We then tasked 13 generative models with predicting the political position of the dialogue locution and proposition stored within each node of the argumentation graph. From this, we were able to compute an ensemble average political position and show how the variance in those predictions was reduced by removing smaller large language models (LLMs). Results demonstrate that the utterances and resolved propositions were, on average, estimated to be left of centre, with the average political position per episode changing, possibly reflecting different locations where the television programme took place within the UK. The argumentation framework is stored within an open graph database management system so that it can be used for graph retrieval-augmented generation (RAG) of UK political personas.

1 Introduction

Before the recent advancements in LLMs, techniques like BertScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021) and GPTScore (Fu et al., 2023) were employed to evaluate outputs from natural language generation tasks. Since then, powerful LLMs have been employed as judges that evaluate the outputs from other LLMs (Chen et al., 2023; Zhang et al., 2023; Chen et al., 2024b; Wang et al., 2024; Chen et al., 2024a). LLMs have demonstrated remarkable performance in following instructions, answering questions and reasoning tasks (Huang and Chang, 2023; Zhao et al., 2025).

Language models have been employed for prediction tasks in the political science literature. Wu et al. (2023) used LLMs to predict the political leaning of a number of politicians using different ideological axes, such as gun control. Ornstein et al. (2025) asked GPT-3 and GPT-4 to predict the probability of a sentence being either *conservative* or *liberal* across a political party manifesto, taking the sentence-level average as an analogue for the political position of said manifestos. Le Mens and Gallego (2025) tasked a variety of LLMs with predicting political leanings of sentences contained within sets of Tweets, British party manifestos, and EU policy speeches in ten different languages.

In this paper, we go beyond the work of Le Mens and Gallego (2025) to provide a computational argumentation-based approach by asking LLMs to predict the political position of locutions and their resolved propositions across 30 debates that were annotated to identify their arguments. The annotated corpora were stored in the Argument Interchange Format (AIF) (Argumentation Research Group, 2011) which we translated to an ASPIC$^+$ argumentation theory (Prakken, 2010; Modgil and Prakken, 2014) and then to a Dung argumentation framework (Dung, 1995), which in turn was instantiated within a Neo4j (2012) graph, a graph database management system (GDBMS). Our experiments demonstrate that LLMs are able to identify political content within locutions and resolved propositions, commonly referred to as argumentative discourse units (ADUs) (Peldszus and Stede, 2013), such that models are capable of making claims about the political leanings of the atoms of arguments. However, whether model predictions are accurately calibrated with the ideologies of the political left and right is a currently unanswered question. Our results build up evidence that LLMs can function as political prediction agents, whilst also showing that language models can complete such tasks when given guidance on argumentation structures revealed within debates.

Our key contributions and findings are:

- We make available an open-source knowledge base containing arguments, relations, and an

ensemble of political position prediction results from 30 televised, political debates. The graph is intended to be used within RAG systems to aid in the emulation of UK political personas.

- We show that LLMs can employ dialogue locutions and their corresponding, resolved propositions in the prediction of political positions.
- Our results show that the political position identified across the 30 debates was, on average, left of centre.
- Removing smaller LLMs from the ensemble average resulted in a decrease in the variance of model predictions.

2 Method

In this section, we describe the methodology for the instantiation of a knowledge base and how we prompted a number of LLMs to predict political leanings. We also list the models and parameters employed in the tests presented later in this paper.

2.1 Graph Instantiation

There is a wealth of open data annotated for arguments on AIFdb.[1] The GDBMS presented in this paper was instantiated using the *QT 30* dataset[2] (Hautli-Janisz et al., 2022) which contains annotated data from 30 debates that featured on the BBC's Question Time (QT). Tab. 3 (in App. D) presents the dates of each debate. The dataset possesses a wide range of political positions on a variety of topics. The dataset is the world's largest annotated corpora that was instantiated using Inference Anchoring Theory (IAT) (Budzynska et al., 2014), an argument annotation scheme that combines speech act theory with argumentation. The data was annotated for utterances (or locutions) which were resolved by annotators into propositions containing their propositional content such that each proposition is understandable to the lay reader when read in isolation. Propositions were annotated for inferences, conflicts, and rephrases, resulting in an argumentation-based representation of the 30 QT debates. Annotations were performed using OVA+ (Janier et al., 2014), a tool for the visualisation of arguments which stores data in the AIF (Chesñevar et al., 2006; Argumentation Research Group, 2011). We converted the AIF to an ASPIC$^+$ argumentation theory (Prakken, 2010; Modgil and Prakken, 2014) and then to a Dung argumentation framework (Dung, 1995), using the procedure described by Bex et al. (2012) (see App. A for a full description of this process). We restricted the ASPIC$^+$ framework to the set of ordinary premises and defeasible inference rules without preferences. The Dung argument framework was instantiated within a Neo4j (2012) GDBMS comprised of 23,228 nodes and 50,905 edges. Each node possesses the following properties: a dialogue **locution** and its corresponding **proposition**; a **speaker**; the **illocutionary force** and corresponding **utterance type**, according to the IAT schemata; a model and ensemble **array of political position predictions**, the **mean**, the **variance** and **standard deviation**, as well as the number of times that the model predicted a node to be **not applicable (NA)** (explained in Sec. 2.2 below) and the **probability of NA**. An example excerpt of the GDBMS is shown in Fig. 4 (in App. B). All the edges representing relations in the GDBMS are directed and comprise the following: **TRANSITIONS_TO** which provides an indication of the chronology of utterances; **IS_A_REPHRASE_OF** which describes when one node is a rephrase of another; **SUPPORTS** which denotes an inference between two nodes; and **ATTACKS** which refers to a conflict between two nodes.

2.2 Populating the Graph with Political Positions

The work presented in this paper was inspired by Le Mens and Gallego (2025) and we extend it by including both a locution and proposition within the prompt provided to the LLMs. Both locutions and propositions are output as a result of the IAT annotation process. Models were tasked with scoring each node's locution and proposition using a scale from 0 to 100, where 0 denoted an *extremely left*-wing view and 100 an *extremely right*-wing stance. In cases where a node's locution and proposition contained no political leaning, models were asked to score such nodes as NA. An example prompt template is provided in App. C.

Given that LLMs are non-deterministic, even when using zero temperature, the same seed and a small top_p, we repeated each prompt five times per model, thus obtaining five potentially different predictions about the political position of a locution and corresponding proposition for each node within the Neo4j (2012) graph. We employed a

[1] https://corpora.aifdb.org/
[2] https://corpora.aifdb.org/qt30

software package, called *Golem* (Blackwell, 2024), to prompt LLMs. The models employed and their parameters are displayed in Tab. 4 (in App. E).

3 Results

We now present our main results, with additional results included in App. F. Each node in the graph contains five predictions from each model, as well as the mean, variance, standard deviation, number of NA scores and the probability that a node was labelled NA, for each model. Each node also contains two ensemble arrays containing the political predictions of all and some models, from which the mean, variance, standard deviation, and NA count and probability was computed and stored.

3.1 Models Included in the Ensembles

We chose how to group model predictions into ensemble results systematically. The first ensemble array was just the set of all model predictions for each node in the knowledge base. The second ensemble was formed by comparing the distribution of political position counts and NA prediction counts for each model (see Fig. 5 in App. F). If the NA count was below the highest (binned) political position count, then, taking the view that such models are too conservative in attributing political opinions to the given inputs, that model's predictions were removed from the second ensemble. It came to light that only less capable models, with fewer parameters, were the ones that were not included in the second ensemble, which we attribute to smaller models' inability to distinguish between apolitical propositions and locutions. For example, consider the distributions of Claude 3.5 Haiku, GPT 4o Mini, Llama 3.1:8b, and Mistral:7b in Fig. 5 (in App. F) where those models score the political position of a node as 50 when they were unsure, instead of scoring the node as NA. Moreover, in Fig. 5, Llama 3.2:3b, the smallest model, further confirms this as it scored the fewest number of nodes NA, whilst predicting that the vast majority of nodes were *extremely left wing*. Model to model comparisons in Fig. 6 further confirmed that smaller models' predictions were tantamount to random, justifying their removal from the second ensemble. The models[3] used for the prediction of political positions and the two ensembles described above are listed in Tab. 1.

Ensemble	Models Included
Ensemble 1	Claude 3.5 Haiku, Claude 3.7 Sonnet, GPT 3.5 Turbo, GPT 4o, GPT 4 Turbo, GPT 4o Mini, GPT o3 Mini, Llama 3.1:8b, Llama 3.2:3b, Mistral:7b, DeepSeek-V3, Gemini 1.5 Pro, and Grok 2.
Ensemble 2	Claude 3.7 Sonnet, GPT 3.5 Turbo, GPT 4o, GPT 4 Turbo, GPT o3 Mini, DeepSeek-V3, Gemini 1.5 Pro, and Grok 2.

Table 1: The models for which political position results were obtained and the two ensembles.

3.2 Distribution of Political Positions

Models' predictions across all nodes and episodes were, on average, left of centre (Fig. 1 and Tab. 2). The removal of smaller models from the ensemble resulted in a reduction in variance and a mean that was closer to 50, as per Tab. 2.

Ensemble	Mean	Median	Standard Deviation, σ
Ensemble 1	42.3	50.0	21.9
Ensemble 2	44.1	50.0	19.0

Table 2: Both ensembles' mean, median and standard deviation.

3.3 Political Position Over Time

Model predictions were, on average, left of centre across all episodes (Fig. 2). The ensemble mean political position ($\pm\sigma$) plotted over time further demonstrates that the removal of less capable models reduced variance in model predictions (Fig. 2).

4 Discussion and Conclusions

Our results show that LLMs can employ argumentation data in the prediction of political positions, and this paper is a first attempt to make use of argumentation data for this task. We observed changes in the mean political position per episode that may correlate with different panellists' and audience members' views on specific topics featured on the programme on different dates, and also filming locations around the UK. Whilst models were able to predict the political position of nodes, whether those scores agreed with panellist or audience member stances is currently an open question.

LLMs exhibit political biases (Motoki et al., 2024; Rozado, 2024; Agiza et al., 2024; Rettenberger et al., 2025) and model outputs are dependent on each vendor's choice of training data, as well as any post-training methods. Our results highlight this notion as the distribution of political positions differs when considering the outputs from different base models and publishers. For instance,

[3]While we tried to obtain results for OpenAI's o1 model, we were not able to collect results for every prompt using o1 because of content filtering, even though all custom content filters were turned off during experiments.

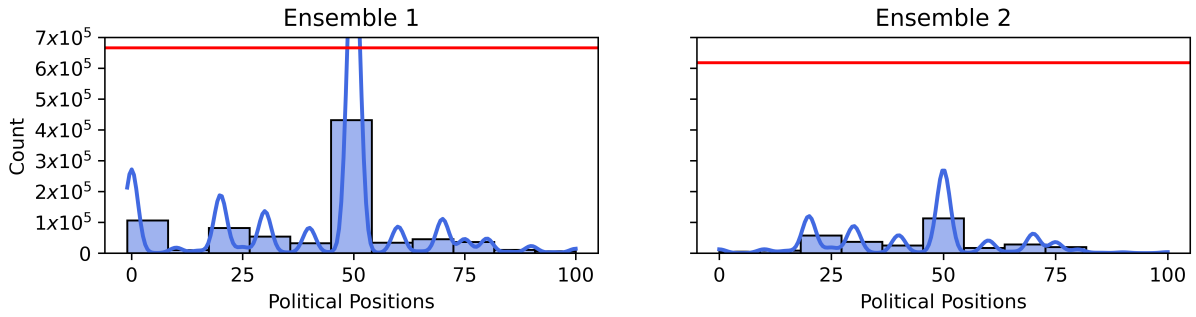

Figure 1: Distribution counts of political positions for all models included (left) and less capable/smaller models removed (right). The NA counts are plotted in red.

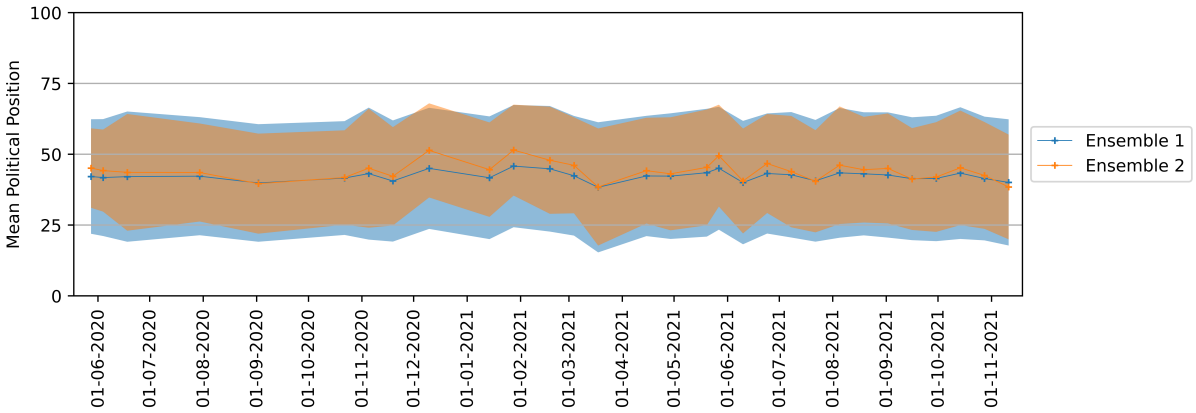

Figure 2: Both ensembles' mean political position ($\pm\sigma$) plotted against time, across all 30 episodes.

in Fig. 5 (in App. F), the distribution of political leanings for DeepSeek-V3 is different to Grok 2 which is different to OpenAI's models, etc. Since politics and training data bias may differ between LLM developers, we advocate using an ensemble average, as we have done in our experiments.

At the node-level, predictions were not deterministic, even for zero temperature, the same seed and low top_p, which was to be expected and can be attributed to the stochastic nature of LLMs. Frequency of models scoring a node as NA was greater than 6×10^5 for both Ensemble 1 and 2 in Fig. 1. Less capable models – such as Llama 3.1:8b, Llama 3.2:3b, Mistral:7b within Ensemble 1 – did not perform well at identifying locutions and propositions that were NA, when compared with the outputs from models in Ensemble 2. However, in reality, there was a negligible difference between NA counts in Ensemble 1 and 2 in Fig. 1. We believe that the method of text segmentation within the *QT 30* dataset may have affected models' ability to differentiate between nodes, containing locutions and propositions, on a political side and those that were non-partisan.

The GDBMS makes it simple to extract pertinent UK political data on a wide range of topics, such as vaccination, immigration, etc. Our knowledge base is both novel and useful for graph RAG for two reasons. First, the GDBMS is a representation of 30 real-world debates where inference, conflict, rephrase, and transition relations between nodes was instantiated (see Section 2.1). So, if one wanted to provide a conversational LLM with pertinent, argumentative examples of, say, conflict, then this is made possible with our GDBMS. Second, the graph's node properties (i.e., the locution, proposition, speaker, political position results, etc) combined with those relations should provide researchers with a way to create UK political personas, using graph RAG, to reflect political leanings. Researchers can choose to use the ensemble or individual models' political position results. Furthermore, notions of uncertainty could be incorporated into future graph RAG systems through use of the probability of a node being NA. A link to the GDBMS will be provided *here* if the paper is accepted (see App. G for a discussion of the licensing arrangements).

Limitations

Ground Truth. We did not check whether model outputs were in agreement with the political positions of panellists. While it might not be possible to ascertain the political stance of audience members, one could easily discover the political leaning of certain panellists, especially if they were a politician. Future work should study whether model outputs are in accordance with the stated political stance of panellists.

Political viewpoint calibration. Interpretation of the scoring scale, between 0 and 100, by the LLMs is necessarily subjective. While we have assumed that models can determine that, say, a score of 50 would indicate a political stance in the centre, models' perceived centre reference point might be different. Future work could address this by switching the political scale so that 0 denotes an *extremely right-wing* and 100 an *extremely left-wing* dialogue locution and proposition to discover whether the resulting outputs are a mirror reflection of the original scores, or whether the set of nodes with a score of, say, 50 changed. In addition, given that we do not have an absolute scale of political positions for models, one idea for future work might be to ask models to describe the ideals associated with integer values on the political scale employed in this paper. Alternatively, we could compare models' outputs by considering the relative ordering on the 0 to 100 scale of pairs of locutions and propositions. An overall comparison could then be computed using a summed edit distance. Such a measure would be independent of values on the absolute scale but rather compare relative political positions of different LLMs.

Text Segmentation. We acknowledge that our experiments employed locutions and propositions (or ADUs), which were usually clauses within the *QT 30* dataset. While propositions are the atoms of arguments, whether performance would increase when providing models with more than just one clause is a task left for future work. We also note that comparing the complexity of text with a prediction about its political stance is interesting, but the clause segmentation approach used in the *QT 30* dataset meant that this comparison was not possible.

Model Imbalance. We note that the ensembles presented contain more models from OpenAI than any other vendor which could have inadvertently impacted the ensemble averages. Future work should consider creating ensembles that only include predictions from one model per vendor to understand whether the distribution of political positions changes as a result.

Coarse-grained analysis. While the main contribution of this paper is our open-source knowledge base, it is possible to derive more results from our graph. For instance, it may be possible to compare the mean political position across different topics that feature in the QT debates, such as immigration, lockdowns, vaccinations, etc. Another interesting point of analytical interest could be how panellists', audience members' and moderator's political leanings evolved over time. Moreover, we have not yet evaluated the role that stochasticity played in our results and, as such, future work could look at the average standard deviation across nodes for each model in our dataset. Finally, a participant study should be conducted to quantify the accuracy of model predictions.

Acknowledgments

Acknowledgement of funding sources omitted for blind review and to be added on acceptance.

References

Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):2–12.

University of Dundee Argumentation Research Group, School of Computing. 2011. The argument interchange format (aif) specification. Retrieved from: https://www.arg-tech.org/wp-content/uploads/2011/09/aif-spec.pdf. (Accessed on: 25.06.2024).

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.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Computational Models of Argument*, Frontiers in artificial intelligence and applications, pages 185–196, Netherlands. IOS Press. Fifth International Conference on Computational Models

of Argument, COMMA 2014 ; Conference date: 09-09-2014 Through 12-09-2014.

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.

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.

C. Chesñevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S Willmott. 2006. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(4):293–316.

Phan M. Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

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 (ELRA). © European Language Resources Association (ELRA).

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.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. Ova+: an argument analysis interface. In *Computational Models of Argument*, Frontiers in artificial intelligence and applications, pages 463–464, Netherlands. IOS Press. Fifth International Conference on Computational Models of Argument, COMMA 2014 ; Conference date: 09-09-2014 Through 12-09-2014.

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

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

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Neo4j. 2012. Neo4j - the world's leading graph database.

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.

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.

I. Rahwan and C. Reed. 2009. *The Argument Interchange Format*, pages 383–402. Springer US, Boston, MA.

I. Rahwan, F. Zablith, and C. Reed. 2007. Laying the foundations for a world wide argument web. *Artificial Intelligence*, 171(10):897–921. Argumentation in Artificial Intelligence.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):1–17.

David Rozado. 2024. The political preferences of llms. *PLOS ONE*, 19(7):1–15.

D. Walton, C. Reed, and F. Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, New York.

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. *Preprint*, arXiv:2303.12057.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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.

A Translating AIF to ASPIC⁺ to a Dung AF

In this appendix, we explain how to tranform data stored in the AIF to an ASPIC⁺ argumentation theory and then to a Dung argumentation framework.

A.1 Abstract Argumentation Frameworks

Dung's seminal work on abstract argumentation frameworks made a significant contribution to the field of computational argumentation and non-monotonic reasoning (Dung, 1995). The underlying notion of his proposal was that arguments and attacks between them can be modelled using a directed graph, where the arguments and attacks are represented as nodes and edges, respectively. Below we provide the formal definition for Dung's original abstract argumentation framework.

Definition A.1.1 A (finite) Dung argumentation framework $\mathcal{G}$ is a tuple $(\mathcal{A}, \mathcal{R})$ which contains a set of arguments $\mathcal{A}$ and binary attack relations $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ between arguments. For two arguments $a_1, a_2 \in \mathcal{A}$, the argument a_1 attacks a_2 if and only if $(a_1, a_2) \in \mathcal{R}$.

We do not consider the acceptability of arguments in this work so we omit the introduction of the semantics.

A.2 The ASPIC⁺ Framework

While Dung's seminal account of abstract argumentation allows for the identification of sets of admissible arguments, its level of abstraction means it pays no attention to the internal structure of those arguments. The ASPIC⁺ framework, a structured argumentation system, adopts an intermediate level of abstraction to provide an abstract account of real-world arguments (Prakken, 2010; Modgil and Prakken, 2014). The ASPIC⁺ framework can model structured argumentation problems using strict and defeasible inference rules, with preferences between defeasible rules, as well as a knowledge base which contains sets of necessary axioms (or facts), ordinary premises, and assumptions. However, we restrict the ASPIC⁺ framework so that it is only comprised of a knowledge base containing the set of ordinary premises and defeasible inference rules without preferences, as they are the only conditions that pertain to the work presented throughout this paper.

Definition A.2.1 (Prakken, 2010). An *argumentation system* is tuple $AS = (\mathcal{L}, ^-, \mathcal{R})$ where

- $\mathcal{L}$ is a logical language;
- $^-$ is a contrariness function $^- : \mathcal{L} \mapsto 2^{\mathcal{L}}$; and
- $\mathcal{R}$ is a set of defeasible rules.

Definition A.2.2 (Prakken, 2010) Let $\mathcal{L}$ be a logical language and $^-$ be a contrariness function, where ϕ and ψ are statements within that language, such that $\phi, \psi \in \mathcal{L}$, and $\overline{\phi}$ and $\overline{\psi}$ are sets containing statements that conflict with ϕ and ψ, respectively. It follows that

- ϕ is called a *contrary* of ψ if and only if $\phi \in \overline{\psi}$;
- ϕ and ψ are *contradictory* if and only if $\psi \in \overline{\phi}$ and $\phi \in \overline{\psi}$, denoted by $\phi = -\psi$.

The arguments formed using the ASPIC⁺ framework are defined as inference trees which are created by applying defeasible inference rules on objects within the logical language. We refer to $p \Rightarrow q$ as a defeasible rule $r \in \mathcal{R}$, where p is the antecedent and q is the consequent.

Definition A.2.3 (Prakken, 2010) For the restricted version of an ASPIC⁺ argumentation system $(\mathcal{L}, ^-, \mathcal{R})$ presented, a *knowledge base* $\mathcal{K}$ is a set of *ordinary* premises $\mathcal{K}_p$, such that $\mathcal{K} \subseteq \mathcal{L}$ and $\mathcal{K} = \mathcal{K}_p$.

Definition A.2.4 (Prakken, 2010) An *argumentation theory* is a pair $AT = (\mathcal{K}, AS)$ where $\mathcal{K}$ is a knowledge base containing the set of ordinary premises $\mathcal{K}_p$ only, such that $\mathcal{K} = \mathcal{K}_p$, and AS is an argumentation system.

Arguments are derived from the knowledge base, where each argument A is obtained from the set of ordinary premises within the knowledge base $\mathcal{K}_p$ of an argumentation theory AT; $Prem(A)$ returns all the ordinary premises within $\mathcal{K}_p$ which support A, $Conc(A)$ returns A's conclusion, and $Sub(A)$ returns all of A's sub-arguments.

Definition A.2.5 (Prakken, 2010) Let $\mathcal{K}$ be a knowledge base in an argumentation system $(\mathcal{L}, ^-, \mathcal{R})$. An argument A is defined as

- $A = \{\phi\}$ if and only if $\phi \in \mathcal{K}$ where $Prem(A) = \{\phi\}$, $Conc(A) = \{\phi\}$, and $Sub(A) = \{\phi\}$.
- $A = \{A_1, ..., A_n \Rightarrow \psi\}$ if and only if $A_1, ..., A_n$ are arguments and there exists a defeasible rule in the argumentation system AS such that $Conc(A_1), ..., Conc(A_n) \Rightarrow \psi \in \mathcal{R}$; $Prem(A) = Prem(A_1) \cup ... \cup Prem(A_n)$; $Conc(A) = \psi$; and $Sub(A) = Sub(A_1) \cup ... \cup Sub(A_n) \cup \{A\}$.

Attacks from one argument to another are represented through the contrariness function, and successful attacks are defined as defeats. The ASPIC$^+$ framework can model *rebutting*, *undermining*, and *undercutting* attacks. A *rebutting* attack is one where an argument attacks the conclusion of another. An *undermining* attack is one where the conclusion of an argument is contrary to the premise of another. An *undercutting* attack is one where an argument's defeasible inference from a set of premises to a conclusion is attacked by another argument. While both rebutting and undercutting attacks can only feature in structured argumentation that allows for defeasible inference rules, undercutting attacks are not included in the new work presented in this paper. Furthermore, as the ASPIC$^+$ formalism has been restricted to the set of defeasible rules without preferences, all conflicts that feature in the sets output by the contrariness function are deemed to be defeats, as per in Definition A.2.6.

Definition A.2.6 (Prakken, 2010) For the arguments $A, B \in \mathcal{K}$, the argument A *defeats* B, when:

- A *rebuts* B (on B') if and only if $Conc(A) \in \overline{\psi}$ for some $B' \in Sub(A)$ where $B' = \{B''_1, ..., B''_n \Rightarrow \psi\}$;
- A *undermines* B (on ψ) if and only if $Conc(A) \in \overline{\psi}$ for some $B' = \psi$, $\psi \in Prem(B)$.

As in (Prakken, 2010), structured argumentation theories along with the set of defeats can be employed in the instantiation of Dung abstract argumentation graphs. Remembering that a Dung-style argument system is a tuple $(\mathcal{A}, \mathcal{R})$ with a set of arguments $\mathcal{A}$ and defeats $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$, as defined in Section A.1.

Definition A.2.7 (Prakken, 2010) A Dung argumentation framework $\mathcal{G}_{AT}$ corresponding to an argumentation theory AT is a pair $(\mathcal{A}, \mathcal{R})$ with a set of arguments $\mathcal{A}$ and relations $\mathcal{R}$ instantiated using the respective arguments (Definition A.2.5) and defeats (Definition A.2.6) within the theory.

A.3 The Argument Interchange Format

The AIF is a community-led attempt to gather a variety of types of works within the computational argumentation literature by providing a shared ontology to facilitate future research and development of argumentation-based tools and techniques (Chesñevar et al., 2006). The onotology acts as an abstract medium that allows researchers to employ any logical language they so choose in order to create argument systems, whilst also providing them with the added benefit of going between languages and formalisms. For instance, if a user were to conduct argument analysis on a debate using an annotation scheme, such as the IAT (Budzynska et al., 2014), and save their data in the AIF, then they would also be able to semantically evaluate the acceptability of arguments in their analysis by mapping it from the AIF to an ASPIC$^+$ argumentation theory and then to a Dung argumentation framework, which can be evaluated using all the well-know semantics. As such, the AIF provides a solid foundation on which real-world applications for argumentation can be based.

The specification for the AIF ontology is presented in Fig. 3. The AIF ontology has two parts, namely the *Upper Ontology* and the *Forms Ontology* (Rahwan et al., 2007; Rahwan and Reed, 2009). The *Upper Ontology* is comprised of *information nodes* (*I-nodes*) and *scheme nodes* (*S-nodes*), allowing users to build the nodes and edges found within argument systems. Depending on the context, *information nodes* store the data points within an argument analysis, such as locutions and propositions, whilst *scheme nodes* capture general patterns of reasoning, such as the inference between a set of premises supporting a conclusion or a conflict between I-nodes. As such, scheme nodes are the instantiation of: rule-application nodes (RA-nodes), indicating an inference from at least one I-node to another; conflict-application nodes (CA-nodes), indicating a conflict between two I-nodes; or preference-application nodes (PA-nodes), which annotate preferences between I-nodes.

The *Forms Onotology* employs the nodes and

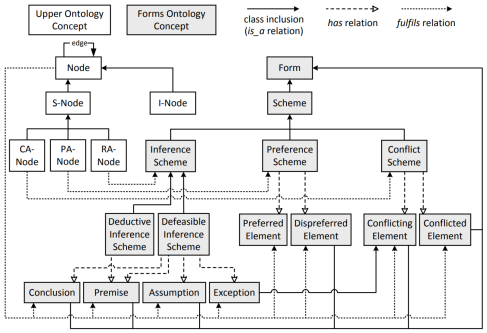

Figure 3: The AIF specification [taken from (Argumentation Research Group, 2011)].

edges in the Upper Ontology and allows users to refine the simple patterns of reasoning captured, instantiating different theoretical argumentative forms, such as argumentation schemes (Walton et al., 2008), thus allowing users to attain a better understanding of argumentation by modelling the intricacies found within it.

The AIF ontology's main representational language is a directed graph. Graphs provide a structured and systematic way of describing argumentation without the constraints of a logic (Chesñevar et al., 2006), while also aligning with many of the accounts of argumentation proposed within the literature. An AIF argument graph $\mathcal{G}_{AIF}$, not to be confused with the AIF ontology specification presented in Fig. 3, is defined in Definition A.3.1.

Definition A.3.1 Let $\mathcal{G}_{AIF} = (V, E)$ be an AIF argument directed graph which is a pair (V, E) where

1. $V = I \cup RA \cup CA$ is the set of vertices in $\mathcal{G}_{AIF}$, where I are the I-nodes, RA are the RA-nodes, and CA are the CA-nodes;
2. $E \subseteq V \times V \setminus I \times I$ is the set of edges in $\mathcal{G}_{AIF}$;
3. if and only if $v \in V \setminus I$, then v has at least one direct predecessor and successor;
4. if and only if $v \in RA$, then v has at least one predecessor and successor in the form of a *premise* and *conclusion*, respectively;
5. if and only if $v \in PA$, then v has exactly one predecessor v_i and one direct successor v_j that instantiates the form *preferred* and *dispreferred element*, respectively, where $v_i \neq v_j$; and
6. if and only if $v \in CA$, then v has exactly one predecessor and successor, respectively termed *conflicting* and *conflicted elements*.

A.4 Translating from the AIF to ASPIC$^+$

After introducing a restricted case of the ASPIC$^+$ framework (Section A.2), which can be employed in the instantiation of a Dung argumentation framework (Definition A.2.7), presenting the AIF (Section A.3), and defining how to instantiate AIF argument graphs (Definition A.3.1), we now present a method proposed by Bex et al. (2012) which is capable of mapping argument data stored in the AIF to an ASPIC$^+$ argumentation theory and then to a Dung argument system, enabling the instantiation of the Neo4j (2012) graph presented in this paper.

Definition A.4.1 For an AIF directed argument graph $\mathcal{G}_{AIF} = (V, E)$ with a set of I-nodes, forms $\mathcal{F}$ and a set of relations that link the nodes in $\mathcal{G}$ to nodes in $\mathcal{F}$, an ASPIC$^+$ argumentation theory AT based on $\mathcal{G}_{AIF}$ is defined as follows:

1. $\mathcal{L} = I \cup RA$;
2. $\mathcal{K}_p = \{v \in I \mid v \text{ is an initial node}\}$ where $v \in \mathcal{K}_p$;
3. $\mathcal{R}$ is the smallest set of inference rules $r_k : v_1, ..., v_n \Rightarrow v$ where there exists a node $v_i \in RA$ such that:
   (a) v_i fulfils a *defeasible* scheme in the Forms Ontology $\mathcal{F}$; and
   (b) v_i's direct predecessors are *premises* of the form $v_1, ..., v_n$ and v_i's direct successor is a conclusion of the form v.
4. $v_h \in \overline{v}_j$ if and only if $\exists v_i \in CA$ such that v_i has a direct predecessor v_h and direct successor v_j.

B Neo4j (2012) Bloom GUI

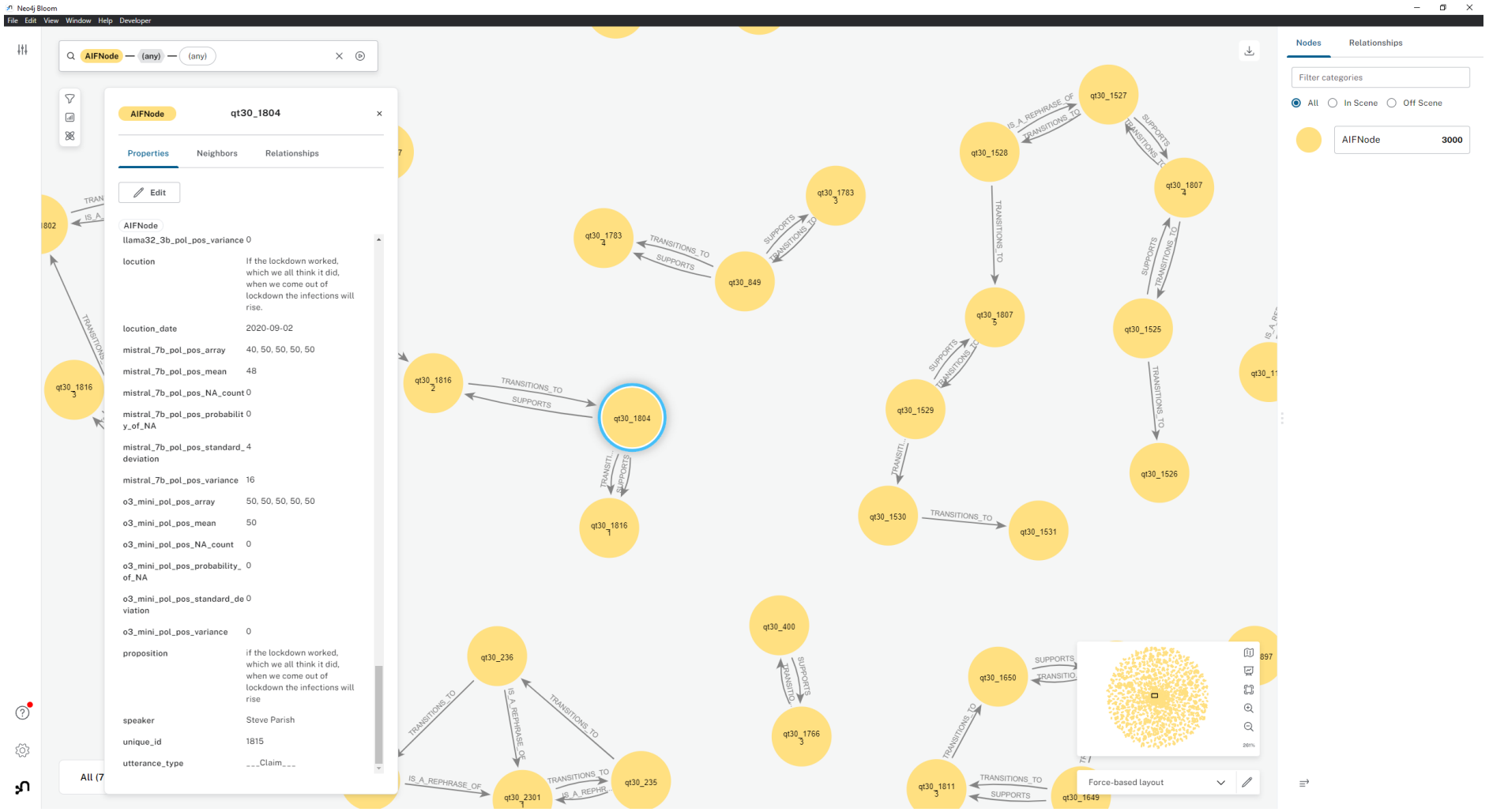

Figure 4: A screenshot of the Bloom GUI (Neo4j, 2012).

C Detail of Prompt Generation

prompt

```
{
   "role": "system",
   "content": "You will be provided with the
   text of a locution and its corresponding
   propositional content that forms part of
   an argument from a UK political debating TV
   programme.
   Your task is to decide where does the
   speaker stand on the 'left' to 'right' wing
   scale using the speaker's locution and
   propositional content?
   Provide your response as a score between 0
   and 100 where 0 means 'Extremely left' and
   100 means 'Extremely right'. If the text
   does not have political content, set the
   score to 'NA'.
   Output in JSON format using the following
   template: {'Score' : int}.
   Do not include any additional context,
   preamble, or explanation."
},
{
   "role": "user",
   "content": "Score the speaker's political
   position between 0 ('Extremely left')
   and 100 ('Extremely right'), and 'NA' if
   there is no political content, using the
   following locution and proposition.
     Proposition: '[Insert Proposition]'
     Locution: '[Insert Locution]'
     Do not write an introduction or summary.
   Output in JSON format using the following
   template: {'Score' : int}"
}
```

D Dates of QT Episodes

Episode	Date
1	28 May 2020
2	4 June 2020
3	18 June 2020
4	30 July 2020
5	2 September 2020
6	22 October 2020
7	5 November 2020
8	19 November 2020
9	10 December 2020
10	14 January 2021
11	28 January 2021
12	18 February 2021
13	4 March 2021
14	18 March 2021
15	15 April 2021
16	29 April 2021
17	20 May 2021
18	27 May 2021
19	10 June 2021
20	24 June 2021
21	8 July 2021
22	22 July 2021
23	5 August 2021
24	19 August 2021
25	2 September 2021
26	16 September 2021
27	30 September 2021
28	14 October 2021
29	28 October 2021
30	11 November 2021

Table 3: The dates of the debates within *QT 30*.

E Models and Parameters

Model	Temperature	Seed	top_p	Weights	Execution	Publisher	Hosted By
Claude 3.5 Haiku (`claude-3-5-haiku-20241022`)	0	NA	0.1	Closed	API	Anthropic	Anthropic
Claude 3.7 Sonnet (`claude-3-7-sonnet-20250219`)	0	NA	0.1	Closed	API	Anthropic	Anthropic
GPT 3.5 Turbo (`gpt-3.5-turbo-0125`)	0	123	0.1	Closed	API	OpenAI	Azure
GPT 4o (`gpt-4o-2024-08-06`)	0	123	0.1	Closed	API	OpenAI	Azure
GPT 4 Turbo (`gpt-4-turbo-2024-04-09`)	0	123	0.1	Closed	API	OpenAI	Azure
GPT 4o Mini (`gpt-4o-mini-2024-07-18`)	0	123	0.1	Closed	API	OpenAI	Azure
GPT o3 Mini (`o3-mini-2025-01-31`)	NA	123	NA	Closed	API	OpenAI	Azure
Llama 3.1:8b (`llama3.1:8b`)	0	123	0.1	Open	Local	Meta	NA
Llama 3.2:3b (`llama3.2:3b`)	0	123	0.1	Open	Local	Meta	NA
Mistral:7b (`mistral:7b`)	0	123	0.1	Open	Local	Mistral AI	NA
DeepSeek-V3 (`deepseek-v3`)	0	123	0.1	Open	API	DeepSeek	DeepSeek
Gemini 1.5 Pro (`gemini-1.5-pro-002`)	0	123	0.1	Closed	API	Google	Google
Grok 2 (`grok-2-1212`)	0	123	0.1	Closed	API	xAI	xAI

Table 4: Models, parameters, publishers and hosting.

F Additional Experimental Results

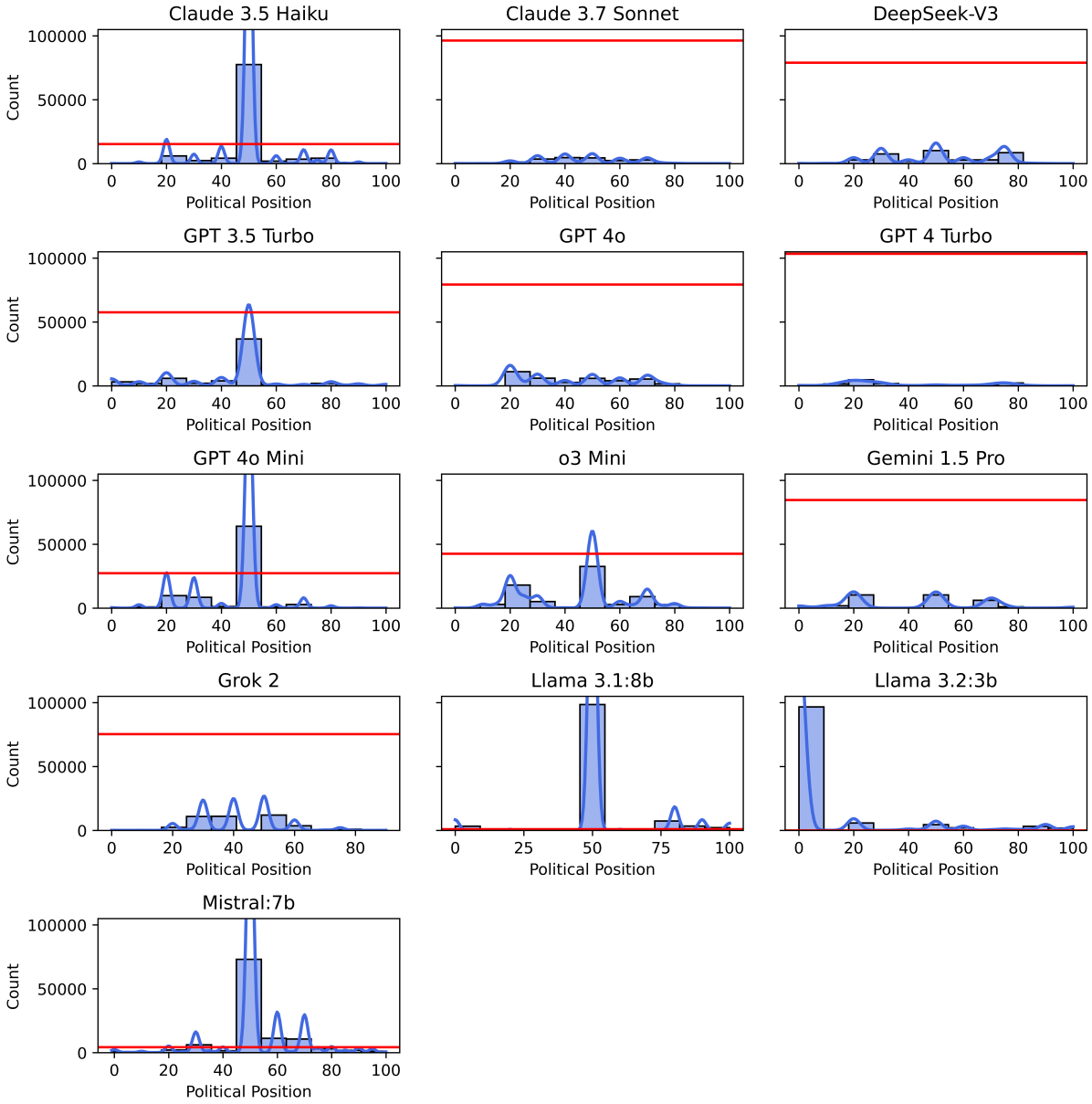

Figure 5: The distribution of political positions (blue) and the NA counts (red) for each model.

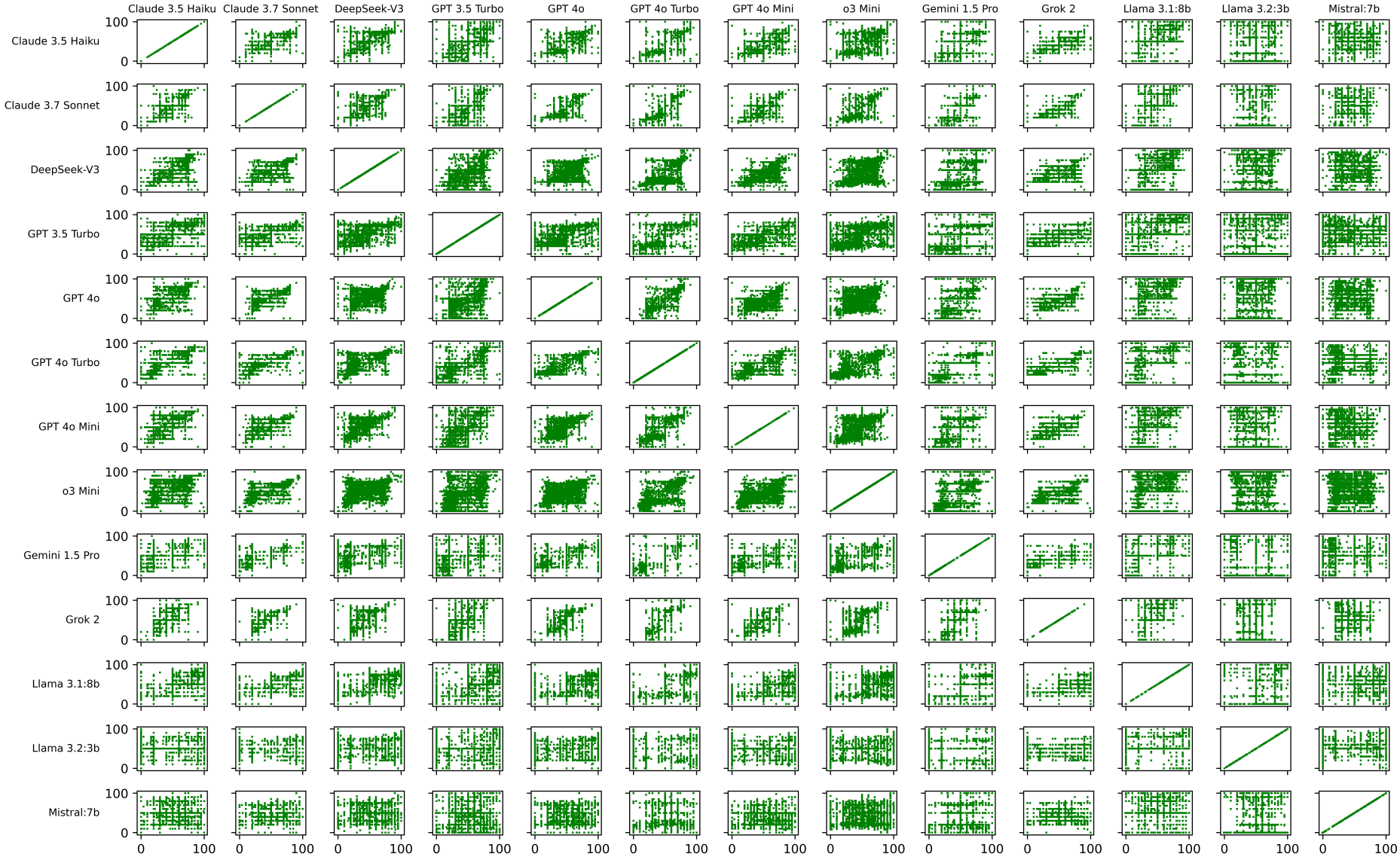

Figure 6: Scatter plots of political position predictions from each model to compare model outputs.

G Licensing

QT 30 is an open-source dataset that was made available without mentioning any specific licensing arrangements – see https://corpora.aifdb.org/qt30. Our use of *QT 30* is consistent with the intended use of the dataset. Our dataset is available under the CC BY 4.0 license.