Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning
KR in the Wild

# Evaluating LLM-Driven Summarisation of Parliamentary Debates with Computational Argumentation

**Eoghan Cunningham**[1] , **James Cross**[2] , **Derek Greene**[1] and **Antonio Rago**[3]
[1]School of Computer Science, University College Dublin
[2]School of Politics and International Relations, University College Dublin
[3]Department of Informatics, King's College London
[1,2]{firstname, surname}@ucd.ie, [3]antonio.rago@kcl.ac.uk

**Abstract**

Understanding how policy is debated and justified in parliament is a fundamental aspect of the democratic process. However, the volume and complexity of such debates mean that outside audiences struggle to engage. Meanwhile, Large Language Models (LLMs) have been shown to enable automated summarisation at scale. While summaries of debates can make parliamentary procedures more accessible, evaluating whether these summaries faithfully communicate argumentative content remains challenging. Existing automated summarisation metrics have been shown to correlate poorly with human judgements of *consistency* (i.e., faithfulness or alignment between summary and source). In this work, we propose a formal framework for evaluating parliamentary debate summaries that grounds argument structures in the contested proposals up for debate. Our novel approach, driven by computational argumentation, focuses the evaluation on argumentative metrics concerning the faithful preservation of the reasoning presented to justify or oppose policy outcomes. We demonstrate our methods using debates from the European Parliament and associated LLM-driven summaries.

## 1 Introduction

Parliamentary debates generate extensive content that must be made comprehensible to citizens, researchers, and policymakers. Debate summaries can make the policy-making process more accessible to those who might not engage with parliamentary activity otherwise (Tsfati 2003), and more understandable for those who do engage (Hwang et al. 2007). Automated summarisation methods are routinely applied to this task (Cunningham, Cross, and Greene 2025), and the introduction of modern Large Language Models (LLMs) has meant that convincing summaries can be generated at scale. However, key challenges remain in evaluating these summarisation methods, particularly assessing faithfulness – sometimes referred to as "factual consistency" (Fabbri et al. 2021). Human evaluations of LLM-driven summaries of argumentative texts have found that they "*lack in dimensions that require logical understanding*" (Li et al. 2024). This is problematic when LLMs are placed between citizens and their representatives. Since such evaluations are not scalable, many studies rely instead on automated measures of lexical overlap or semantic similarity (Li et al. 2024; Roush et al. 2024). These methods often correlate poorly with human judgments of factual consistency (Kryściński et al. 2019). In this paper, we show how formal argumentation frameworks (Baroni et al. 2018) can can help to remedy this issue by providing a structured, interpretable basis for assessing the faithfulness of debate summaries, following existing work in this area (Hautli-Janisz et al. 2022).

We propose an argumentation framework (see §3) to represent the arguments made by politicians to oppose or justify specific policy outcomes; i.e., the key provisions contested in the debate. This serves two purposes. Firstly, it focuses summary evaluations on outcomes, because promoting parliamentary transparency and democratic participation necessitates summaries that faithfully communicate the reasoning presented to justify or undermine these outcomes. Secondly, these root proposals provide the shared anchor points for aligning argumentation frameworks from source debates with those extracted from their summaries. Following this alignment, we evaluate the faithfulness of a debate summary on the basis of five argumentative metrics introduced in §3.

We ground this evaluation setting with a case-study of debates (and associated LLM-driven summaries) from the European Parliament (EP), which provides a proof-of-concept of our approach. Firstly, to demonstrate how these frameworks can be constructed at scale, in §4.1 we benchmark state-of-the-art argument mining methods on an annotated dataset of arguments from EP debates. Secondly, in §4.2, we apply our methods (alongside established metrics from Natural Language Processing (NLP) to a sample of debates to show how our proposed metrics can provide a formal and interpretable basis for evaluating debate summaries.

## 2 Preliminaries

A *quantitative bipolar argumentation framework* (QBAF) (Amgoud and Ben-Naim 2018; Baroni, Rago, and Toni 2018) is a quadruple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ where: $\mathcal{X}$ is a finite set of *arguments*; $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ is a binary relation of (*direct*) *attack*; $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{X}$ is a binary relation of (*direct*) *support*; $\mathcal{A}$ and $\mathcal{S}$ are disjoint; and $\tau : \mathcal{X} \to [0, 1]$ gives the *base score* of an argument. For the remainder of this section we assume as given a generic QBAF $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$. For any argument $x_i \in \mathcal{X}$, we denote $\mathcal{A}(x_i) = \{x_j \in \mathcal{X} | (x_j, x_i) \in \mathcal{A}\}$ as $x_i$'s *attackers* and $\mathcal{S}(x_i) = \{x_j \in \mathcal{X} | (x_j, x_i) \in \mathcal{S}\}$ as $x_i$'s *supporters*. Arguments may be assigned *strengths* representing their acceptability by means of a *gradual semantics* $\sigma$, such that for $x_i \in \mathcal{X}$, $\sigma(\mathcal{Q}, x_i) \in [0, 1]$ is calculated by start-

ing from its initial strength (i.e., its base score), which is then recursively increased towards 1 or decreased towards 0 based on the strengths of its attackers or supporters, respectively. Such gradual semantics satisfy properties that are intuitive for representing debates (de Tarlé, Bonzon, and Maudet 2022; Rago, Li, and Toni 2023). We use the *DF-QuAD semantics* (Rago et al. 2016), whereby for any $x_i \in \mathcal{X}$, $\sigma(\mathcal{Q}, x_i) = c(\tau(x_i), \Sigma(\sigma(\mathcal{Q}, \mathcal{A}(x_i))), \Sigma(\sigma(\mathcal{Q}, \mathcal{S}(x_i))))$ where, for any $S \subseteq \mathcal{X}$, $\sigma(\mathcal{Q}, S)$=$(\sigma(\mathcal{Q}, x_1), \ldots, \sigma(\mathcal{Q}, x_m))$ for $(x_1, \ldots, x_m)$, an arbitrary permutation of $S$. Then, $\Sigma$ is such that $\Sigma(()) = 0$, where $()$ is an empty sequence, and, for $v_1, \ldots, v_n \in [0, 1]$ ($n \geq 1$): if $n = 1$, then $\Sigma((v_1)) = v_1$; if $n = 2$, then $\Sigma((v_1, v_2)) = v_1 + v_2 - v_1 \cdot v_2$; and if $n > 2$, then $\Sigma((v_1, \ldots, v_n)) = \Sigma(\Sigma((v_1, \ldots, v_{n-1})), v_n)$. Finally, $c$ is such that, for $v^0, v^-, v^+ \in [0, 1]$: if $v^- \geq v^+$, then $c(v^0, v^-, v^+) = v^0 - v^0 \cdot |v^+ - v^-|$; and if $v^- < v^+$, then $c(v^0, v^-, v^+) = v^0 + (1 - v^0) \cdot |v^+ - v^-|$. For any $x_i, x_j \in \mathcal{X}$, we let a *path* from $x_i$ to $x_j$ be defined as a sequence $(x_0, x_1), \ldots, (x_{n-1}, x_n)$ for some $n > 0$, where $x_0 = x_i$, $x_n = x_j$ and, for any $1 \leq k \leq n$, $(x_{k-1}, x_k) \in \mathcal{A} \cup \mathcal{S}$. Then, $\mathsf{paths}(\mathcal{Q}, x_i, x_j)$, denotes the set of all paths between any $x_i, x_j \in \mathcal{X}$. With a small abuse of notation, we treat paths as sets of pairs. The *pro* and *con arguments* (Rago, Li, and Toni 2023) of any $x_i \in \mathcal{X}$ are $\mathsf{pro}(\mathcal{Q}, x_i) = \{x_j \in \mathcal{X} | \exists p \in \mathsf{paths}(\mathcal{Q}, x_j, x_i) : |p \cap \mathcal{A}| \text{ is even}\}$ and $\mathsf{con}(\mathcal{Q}, x_i) = \{x_j \in \mathcal{X} | \exists p \in \mathsf{paths}(\mathcal{Q}, x_j, x_i) : |p \cap \mathcal{A}| \text{ is odd}\}$, respectively.

*Bipolar argumentation frameworks* (BAFs) (Cayrol and Lagasquie-Schiex 2005) are triples $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$, i.e., QBAFs without base scores. For any $x_i, x_j \in \mathcal{X}$ such that there is a path from $x_i$ to $x_j$, $(x_1, x_2), \ldots, (x_{n-1}, x_n)$, where $n \geq 3$, this path is: an *indirect attack* from $x_i$ to $x_j$ iff $(x_1, x_2) \in \mathcal{A}$ and $(x_k, x_{k+1}) \in \mathcal{S}$ $\forall k \in \{2, \ldots, n-1\}$; a *supported attack* from $x_i$ on $x_j$ iff $(x_{n-1}, x_n) \in \mathcal{A}$ and $(x_k, x_{k+1}) \in \mathcal{S}$ $\forall k \in \{1, \ldots, n-2\}$; or an *indirect support* from $x_i$ on $x_j$ iff $(x_k, x_{k+1}) \in \mathcal{S}$ $\forall k \in \{1, \ldots, n-1\}$. A set of arguments $X \subseteq \mathcal{X}$, also called an *extension*, is said to *set-attack* any $x_i \in \mathcal{X}$ iff there exists an attack (whether direct, indirect or supported) from some $x_j \in X$ on $x_i$. Meanwhile, $X$ is said to *set-support* any $x_i \in \mathcal{X}$ iff there exists a (direct or indirect) support from some $x_j \in X$ on $x_i$. Then, a set $X \subseteq \mathcal{X}$ *defends* any $x_i \in \mathcal{X}$ iff $\forall x_j \in \mathcal{X}$, if $\{x_j\}$ set-attacks $x_i$ then $\exists x_k \in X$ such that $\{x_k\}$ set-attacks $x_j$. Any set $X \subseteq \mathcal{X}$ is then said to be *conflict-free* iff $\nexists x_i, x_j \in X$ such that $\{x_i\}$ set-attacks $x_j$. The notion of a set $X \subseteq \mathcal{X}$ being *d-admissible* (based on *admissibility* in (Dung 1995)) requires that $X$ is conflict-free and defends all of its elements. Finally, $X$ is *d-preferred* iff it is d-admissible and maximal wrt set-inclusion. We let $\Sigma_d(\mathcal{Q})$ denote the set of all d-preferred extensions in $\mathcal{Q}$ and we use *credulous*, rather than *sceptical*, acceptance of arguments, i.e., to be accepted they must be in one extension, rather than all extensions.

## 3 Formally Representing Parliamentary Debates and Summaries

In this section, we first introduce a formal representation of parliamentary debates, which may be derived from the debate transcipts themselves, or summaries thereof. Then, we introduce argumentative metrics for individual QBAFs and for pairwise comparisons between two QBAFs from the two sources, thus providing a means of assessing summarisation methods formally. We represent the debates as follows.

**Definition 1.** *Given a debate $\mathcal{D}$ on proposal $\mathcal{P}$ containing $n$ provisions, $\mathcal{P} = \{p_1, \ldots, p_n\}$, a* QBAF representing $\mathcal{D}$ *is a QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ such that:*

- $\mathcal{X} = \mathcal{X}_p \cup \mathcal{X}_s$, *where* $\mathcal{X}_p \cap \mathcal{X}_s = \emptyset$ *and:*
  - $\mathcal{X}_p$ *is the set of* proposal arguments *where* $|\mathcal{X}_p| = |\mathcal{P}|$ *and* $\forall p_i \in \mathcal{P}$, *where* $i \in \{1, \ldots, n\}$, $\exists x_i \in \mathcal{X}_p$;
  - $\mathcal{X}_s$ *is the set of* speech arguments;
- $\mathcal{A} \subseteq \mathcal{X}_s \times (\mathcal{X}_s \cup \mathcal{X}_p)$ *and* $\mathcal{S} \subseteq \mathcal{X}_s \times (\mathcal{X}_s \cup \mathcal{X}_p)$;
- $\tau$ *is such that* $\forall x_i \in \mathcal{X}_p$, $\tau(x_i) = 0.5$.

Our QBAFs thus represent all $n$ provisions being debated as proposal arguments, while speech arguments are those put forward by speakers taking part in the debate. Intuitively, attacks and supports come from speech arguments only, and they are directed towards proposal arguments or other speech arguments. Regarding base scores, we fix those of proposal arguments to be the midpoint to allow for the balance between attacking and supporting speech arguments to determine the proposal arguments' strengths. Meanwhile, base scores of speech arguments are left undetermined so that they can be tailored to specific applications, e.g. some may assume a constant base score, while in others, modelling argument similarity may allow the base score to vary using argument prominence, i.e., the number of times the argument was put forward. In our later examples, we assume a base score of $0.2$ for all arguments in $\mathcal{X}_s$, as this gave reasonable results in our preliminary experimentation, and we calculate strength scores using the *DF-QuAD* semantics.[1] We leave experimentation with other (constant or varying) base scores to future work. Examples of our QBAFs are shown in Figure 1 and are assessed along our metrics, which we introduce next, in Table 1.

For our first metric, we consider the relevance of the speech arguments by measuring how many are connected to at least one proposal argument. Speech arguments without a path to the proposal arguments are thus deemed as irrelevant.

**Metric 1.** *The* argument relevance *of any $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ representing a debate $\mathcal{D}$ which concerned proposal $\mathcal{P} = \{p_1, \ldots, p_n\}$ is* $\pi_{ar}(\mathcal{Q}) = \frac{|\{x_i \in \mathcal{X}_s | \exists x_j \in \mathcal{X}_p : \mathsf{paths}(x_j, x_i) \neq \emptyset\}|}{|\mathcal{X}_s|}$.

As shown in Table 1, $\mathcal{Q}$ and $\mathcal{Q}^*_a$ do not have maximal relevance due to the unconnected argument in each, whereas this is not the case for $\mathcal{Q}^*_b$. High relevance may be suitable for some summaries, but in others, maintaining the same relevance as the original debate may be more important, while such "irrelevant" arguments have also been shown to be helpful in argumentative analysis (Ruiz-Dolz et al. 2025).

Next, we consider, within the set of relevant speech arguments, the balance of evidence for and against a given proposal argument. Here, we measure the number of relevant speech arguments *for* a proposal argument, normalised by the sum of those *for* and *against* it.

[1] In the supplementary material (SM) of the technical report for this paper (Cunningham et al. 2026), we report comparable results with other base scores revealing minimal effect on interpretations.

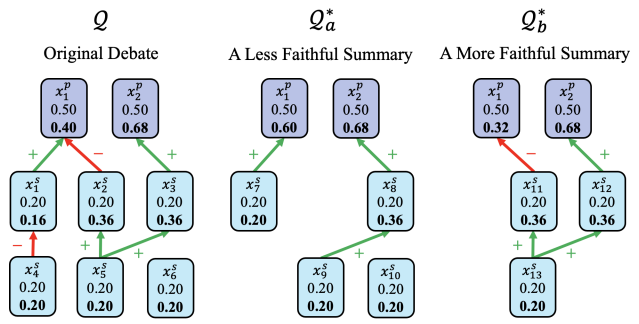


Figure 1: Three example QBAFs representing a debate on two proposals $\mathcal{P} = \{p_1, p_2\}$, where $\mathcal{Q}$ may be taken to be extracted from the original debate text, and $\mathcal{Q}_a^*$ and $\mathcal{Q}_b^*$ from summaries thereof. Proposal (speech) arguments are represented by purple (blue, respectively) nodes with superscript $p$ ($s$, respectively), attacks by red edges labelled "−" and supports by green edges labelled "+". Arguments' base scores (strengths, calculated by the DF-QuAD semantics) are the values in normal (bold, respectively) text.

**Metric 2.** *The* pro-con ratio *of any* $x_i \in \mathcal{X}_p$*, where* $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ *represents a debate* $\mathcal{D}$ *which concerned proposals* $\mathcal{P} = \{p_1, \ldots, p_n\}$*, is* $\pi_{pcr}(\mathcal{Q}, x_i) = \frac{|\mathsf{pro}(\mathcal{Q}, x_i)|}{|\mathsf{pro}(\mathcal{Q}, x_i) \cup \mathsf{con}(\mathcal{Q}, x_i)|}$.

In Table 1, $\mathcal{Q}_b^*$ maintains the same pro-con ratios as $\mathcal{Q}$, which we believe to be desirable in a summary, while $\mathcal{Q}_a^*$ does not due to con arguments of $x_1^p$ not being included.

We now turn to the pairwise metrics. For the remainder of this paper, we assume as given two QBAFs $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ and $\mathcal{Q}^* = \langle \mathcal{X}^*, \mathcal{A}^*, \mathcal{S}^*, \tau^* \rangle$ representing a debate $\mathcal{D}$ concerning proposal $\mathcal{P} = \{p_1, \ldots, p_n\}$. We assume that $\mathcal{Q}$ and $\mathcal{Q}^*$ were extracted from the original debate text and a summarisation thereof, respectively.

We consider pairwise metrics for proposal arguments only. While our framework supports speech argument metrics (see the SM in (Cunningham et al. 2026)), these will require the detection of similar arguments across the two frameworks, which we leave to future work.

Our first pairwise metric checks how many of the proposal arguments' acceptance statuses (by the d-preferred semantics) remained consistent; giving an indication of whether extension-based semantics' results are preserved.

**Metric 3.** *The* d-preferability-consistency *of* $\mathcal{Q}^*$ *wrt* $\mathcal{Q}$ *is* $\pi_{dpc}(\mathcal{Q}^*, \mathcal{Q}) = \frac{|S|}{|\mathcal{X}_p^*|}$*, where* $S \subseteq \mathcal{X}_p^*$ *is the maximal subset such that* $\forall x_i \in S$*,* $\exists X \in \Sigma_d(\mathcal{Q})$ *where* $x_i \in X$ *iff* $\exists X^* \in \Sigma_d(\mathcal{Q}^*)$ *where* $x_i \in X^*$.

Table 1 shows that $\mathcal{Q}_b^*$ preserves the acceptability statuses of both arguments, while $\mathcal{Q}_a^*$ does not for $x_1^p$ since both of its attackers were excluded from this summary.

Our final two metrics concern gradual semantics, so we assume a given $\sigma$. Metric 4 checks whether proposal arguments that received net support (have strength greater than 0.5) or net opposition (have strength less than 0.5) remain so in the debate summary, and is inspired by the *balance* property in (Baroni, Rago, and Toni 2018). This important metric ensures that any summary of a debate remains faithful to, and does not misrepresent, the balance of the discussion, whether in favour of or against the each of the proposals.

| | $\mathcal{Q}$ | $\mathcal{Q}_a^*$ | $\mathcal{Q}_b^*$ |
|---|---|---|---|
| $\lvert\mathcal{X}_s\rvert$ | 6 | 4 | 3 |
| $\pi_{ar}(\cdot)$ | 0.83 | 0.75 | 1.00 |
| $\pi_{pcr}(\cdot, r_1^p)$ | 0.25 | 1.00 | 0.00 |
| $\pi_{pcr}(\cdot, r_2^p)$ | 1.00 | 1.00 | 1.00 |
| $\pi_{dpc}(\cdot, \mathcal{Q})$ | — | 0.50 | 1.00 |
| $\pi_{bc}(\cdot, \mathcal{Q})$ | — | 0.50 | 1.00 |
| $\pi_{\epsilon a}(\cdot, \mathcal{Q})$ | — | 0.50 | 1.00 |

Table 1: The QBAFs from Figure 1 assessed along our argumentative metrics. Note that $\epsilon = 0.1$ was used for $\pi_{\epsilon a}$.

**Metric 4.** *The* balance-consistency *of* $\mathcal{Q}^*$ *wrt* $\mathcal{Q}$ *is* $\pi_{bc}(\mathcal{Q}^*, \mathcal{Q}) = \frac{|S|}{|\mathcal{X}_p^*|}$*, where* $S \subseteq \mathcal{X}_p^*$ *is the maximal subset such that* $\forall x_i \in S$*,* $\sigma(\mathcal{Q}^*, x_i) > 0.5$ *iff* $\sigma(\mathcal{Q}, x_i) > 0.5$*; and* $\sigma(\mathcal{Q}^*, x_i) < 0.5$ *iff* $\sigma(\mathcal{Q}, x_i) < 0.5$.

Table 1 shows that the strengths in $\mathcal{Q}_b^*$ are consistent with those in $\mathcal{Q}$, while, again, $\mathcal{Q}_a^*$ performs poorly in the metric for $x_1^p$ since its strength reflects net support, rather than the net opposition it receives in $\mathcal{Q}$.

Our final metric is a little stricter, checking the number of proposal arguments in the summary that maintain their original strength within a specified error, assessing the accuracy of the summarised proposal arguments' strengths. While we expect some variance in these strengths given that a summary cannot contain all arguments, the error $\epsilon$ should be selected carefully to provide suitable limits.

**Metric 5.** *The* $\epsilon$-accuracy *of* $\mathcal{Q}^*$ *wrt* $\mathcal{Q}$ *is* $\pi_{\epsilon a}(\mathcal{Q}^*, \mathcal{Q}) = \frac{|S|}{|\mathcal{X}_p^*|}$*, where* $S \subseteq \mathcal{X}_p^*$ *is the maximal subset such that* $\forall x_i \in S$*,* $\sigma(\mathcal{Q}^*, x_i) \in [\sigma(\mathcal{Q}, x_i) - \epsilon, \sigma(\mathcal{Q}, x_i) + \epsilon]$.

Table 1 again shows that the strengths in $\mathcal{Q}_b^*$ are within $\epsilon = 0.1$ of those in $\mathcal{Q}$, while $x_1^p$ performs worse for $\mathcal{Q}_a^*$ since its strength here has changed by more than $\epsilon$.

Our running example has thus illustrated that while $\mathcal{Q}_b^*$ gives a more compact summary of $\mathcal{Q}$, by our argumentative metrics it provides a much more faithful summary.

## 4 Can LLMs Summarise European Parliament Debates?

In this section, we show how our approach can be used to evaluate summarisation of parliamentary debates at scale.

We use a dataset of 511 arguments across three EP debates from 2006–2008. Each debate contains between 107 and 198 arguments. The debates follow a structured process where Members of the European Parliament (MEPs) discuss individual policy provisions in a proposal before voting. Each resolution has numbered paragraphs stating provisions. For example, *"The European Parliament recommends that the mandate of the ECDC be extended to non-communicable diseases."* For a given debate, the numbered paragraphs from the debated proposal form the set of proposal arguments $\mathcal{X}_p$. We let the arguments identified in the debate

transcript represent the speech arguments $\mathcal{X}_s$. These arguments, and the attack and support relations between them, are mined from EP debates using the approach detailed next.

### 4.1 Argument Mining

While our framework is agnostic of the method of argument mining, it is important to evaluate the extent to which the framework can be mined reliably and automatically from debates in parliament. For this evaluation, we rely on *relation-based argument mining* (Carstens and Toni 2015). Here, rather than classifying each sentence in a text as argumentative (or non-argumentative) in isolation, we classify all pairs of sentences according to their relation: *support*, *attack*, or *neither*. Thus, we define a sentence as relevant (or *argumentative*) if it attacks or supports an existing argument.

We include a temporal constraint in the mining of $\mathcal{Q}$, such that $(\mathcal{A} \cup \mathcal{S}) \cap (\mathcal{X}_s \times \mathcal{X}_s) \subseteq \{(x_j, x_i) | j > i\}$, where the subscripts of arguments indicate their temporal order in the debate. That is, speech arguments can only attack or support speech arguments that precede them in the debate. Accordingly, $\mathcal{Q}$ contains $n$ multitrees that are built around the root proposals. No such temporal constraint could be assumed for the text of a debate summary; thus, we permit cycles in $\mathcal{Q}^*$, and rely on approximate convergence under the DF-QuAD semantics ($\delta = 10^{-3}$). QBAFs are constructed using Argument Relation Classification (ARC).

To evaluate SOTA models in ARC in this context, we compile a dataset of 481 argument pairs, consisting of 377 arguments. We classify relations into *attack*, *support*, and *neither*, by aggregating annotations across three human coders with expertise in EU politics. Each pair is classified by two coders, and any disagreements are resolved by a third. Cohen's kappa agreement scores between annotators ranged from 'fair' ($\kappa$ = 0.30) to 'moderate' ($\kappa$ = 0.54), indicating that classifying relations is more challenging for some topics than others. While fair-to-moderate agreement scores are typical in argument mining tasks (Lawrence and Reed 2019), an overall agreement of 76.6% indicates that the task involves subjective judgement on the part of the annotators, which may reasonably limit the performance ceiling for automated methods. Following recent studies demonstrating the capabilities of LLMs in (Savigny and Yun 2026; Gorur, Rago, and Toni 2025), we evaluate four different models on this task in various settings (i.e., zero-shot, few-shot, and fine-tuned). We report their performance in Table 2, with results for additional models evaluated (and all prompts used) in the SM in (Cunningham et al. 2026).

Critically, we find that larger reasoning models implemented in zero-shot settings outperform smaller fine-tuned models when applied to new data. This aligns with recent studies of LLMs in argumentative tasks, finding that smaller models struggle even with fine-tuning (Furman et al. 2023), and especially compared to larger models (Bezou-Vrakatseli, Cocarascu, and Modgil 2025). We hope to extend our dataset in future work to investigate further whether LLMs really learn to mine arguments, or just the datasets themselves (Feger, Boland, and Dietze 2025). In a trade-off between cost and performance, we use Claude Haiku to construct the QBAFs in the remainder of our case-study.

| Model | Accuracy | F1 |
|---|---|---|
| Claude Sonnet 4.5 (zero-shot) | **0.74** | **0.78** |
| Claude Haiku 4.5 (zero-shot) | 0.68 | 0.73 |
| Qwen3 30B (zero-shot) | 0.64 | 0.72 |
| Llama 3.1 8B (fine-tuned) | 0.21 | 0.27 |

Table 2: Argument Relation Classification from EP debates. Claude Sonnet and Haiku are proprietary pre-trained LLMs from *Anthropic*. Qwen3 is a open-weight, pre-trained LLM from *Alibaba*. Llama 3.1 (an open-weight model from *Meta*) was fine-tuned on existing 8 ARC datasets by (Savigny and Yun 2026).

| | Debate 1 | | | Debate 2 | | |
|---|---|---|---|---|---|---|
| | $\mathcal{Q}$ | $\mathcal{Q}^*_{\text{Sonnet}}$ | $\mathcal{Q}^*_{\text{Haiku}}$ | $\mathcal{Q}$ | $\mathcal{Q}^*_{\text{Sonnet}}$ | $\mathcal{Q}^*_{\text{Haiku}}$ |
| R-2 | — | 0.24 | 0.05 | — | 0.12 | 0.04 |
| BERT | — | 0.82 | 0.83 | — | 0.84 | 0.84 |
| NLI | — | 0.60 | 0.69 | — | 0.59 | 0.56 |
| $\pi_{ar}(\cdot)$ | 1.00 | 0.97 | 0.93 | 1.00 | 0.97 | 1.00 |
| $\pi_{pcr}(\cdot, r_i^p)$ | 0.53 | 0.52 | 0.58 | 0.50 | 0.52 | 0.65 |
| $\pi_{dpc}(\cdot, \mathcal{Q})$ | — | 1.00 | 1.00 | — | 1.00 | 1.00 |
| $\pi_{bc}(\cdot, \mathcal{Q})$ | — | 0.26 | 0.19 | — | 0.53 | 0.47 |
| $\pi_{\epsilon a}(\cdot, \mathcal{Q})$ | — | 0.52 | 0.30 | — | 0.40 | 0.00 |

Table 3: Comparing LLM-generated summaries to original debate transcripts across two sampled debates. We include traditional NLP-based summarisation metrics (ROUGE, BERTScore, Sentence-NLI) for comparison with our argumentative metrics. For $\pi_{pcr}$, we report the mean over all provisions $r_i^p \in \mathcal{X}_p$.

### 4.2 Summary Evaluation

Next, we generate summaries for a sample of EP debates and apply our proposed methods of summary evaluation alongside three existing metrics. *ROUGE* (Lin 2004) is based on explicit n-gram overlap between the source and summary content. Thus, it is liable to over-penalise fair abstraction or paraphrasing in summaries. *BERTScore* (Zhang et al. 2020) accounts for abstraction by measuring semantic similarity between tokens in the source and the summary. However, when comparing large volumes of text, these methods are unable to sufficiently penalise inconsistencies in summaries in the presence of strong signals of semantic similarity. *Natural Language Inference* (NLI) models are designed to address this issue by detecting factual inconsistencies in the summary via *entailment classification*. However, these metrics evaluate summaries holistically against source text, offering aggregate scores and lacking any formal representation of the argumentation in the debate. As such they obscure which aspects of the debate content are preserved or lost. We list all scores for two sample debates in Table 3 and all prompts used in summarisation are provided in the SM in (Cunningham et al. 2026).

Table 3 shows traditional metrics confirm the summaries are abstractive (low ROUGE), semantically coherent with the source debates (high BERTScore), with many sentences supported by the source (NLI: 0.56–0.69). However, these metrics provide minimal discrimination between summaries

| | $x^p_{10}$ | $x^p_{11}$ | $x^p_{12}$ | $x^p_{13}$ | $x^p_{14}$ | $x^p_{15}$ | $x^p_{16}$ |
|---|---|---|---|---|---|---|---|
| $\mathcal{Q}$ | 0.53 | 0.50 | 0.50 | 0.50 | 0.50 | 0.47 | 0.60 |
| $\mathcal{Q}^*_{\text{Sonnet}}$ | 0.54 | 0.52 | 0.78 | 0.78 | 0.82 | 0.54 | 0.90 |
| $\mathcal{Q}^*_{\text{Haiku}}$ | 1.00 | 0.90 | 0.90 | 0.90 | 0.90 | 0.89 | 0.82 |

Table 4: Strengths for a sample of proposal arguments in QBAFs extracted from the original debate ($\mathcal{Q}$) and LLM summaries ($\mathcal{Q}^*$).

that differ substantially in argument structure. In contrast, our metrics reveal clear degradation patterns. Firstly, Haiku-generated summaries fail to preserve the pro-con ratio of either debate ($\pi_{pcb}$). Second, all summaries struggle to accurately communicate the balance of evidence for and against the key proposals ($\pi_{bc}$ of 0.19–0.53) and largely fail to preserve precise proposal strengths ($\pi_{\epsilon a}$ of 0.00–0.52). Critically, according to the Haiku generated summary, none of the provision strengths are within $\epsilon = 0.1$ of the scores in the original debate ($\pi_{\epsilon a}(\mathcal{Q}^*_{\text{Haiku}}, \mathcal{Q}) = 0$). Meanwhile, $\pi_{dpc}$ gives maximum values only because no proposal arguments were in d-preferred extensions, showing that gradual semantics may be more effective for this task. Unsurprisingly, the larger model (Sonnet) scores better across almost all metrics.

As an illustrative example, Table 4 shows how a subset of the proposal strengths have been misrepresented in the summaries of the second debate. The summaries exclude much of the argumentation that undermined the root provisions. Key proposal arguments that were contested in the debate ($\sigma(\mathcal{Q}, x^p_i) \approx 0.5$) are presented as overwhelmingly supported in the summaries ($\sigma(\mathcal{Q}^*_{\text{Haiku}}, x^p_i) \gg 0.5$).

## 5 Conclusions and Future Work

We proposed a formal argumentation framework for evaluating parliamentary debate summaries that represents debates as QBAFs constructed around contested policy outcomes. The framework demonstrates how computational argumentation can focus evaluations on faithfully communicating the motivations for policy outcomes rather than on general semantic coherence or similarity. We achieve this through argumentative metrics that assess argument acceptability (and thus, the support for proposed outcomes) from the perspectives of both extension-based and gradual semantics, showing that the latter may be more effectively discriminative.

Future work will develop methods for matching similar arguments across sources and summaries to enable speech argument evaluation beyond our current provision-focused metrics. This would support precision-recall analysis of argumentative content preservation, and facilitate identification of speech arguments that are omitted or hallucinated. Further, it would be interesting to assess the usefulness of our metrics via user studies, ensuring that summaries judged by users as being faithful align with our argumentative metrics over standard metrics, and also in downstream tasks. Another fruitful line of work could be to investigate whether QBAFs extracted via LLMs, e.g. as in (Freedman et al. 2025), are faithful according to our metrics. Finally, our approach could be supplemented with existing techniques for argument mining in political debates, e.g. leveraging knowledge graph embeddings (Dore, Faralli, and Villata 2025).

## AI Declaration

The authors did not use generative AI tools in any part of the paper-writing process.

## Acknowledgments

This work was partially supported by Research Ireland grant number SFI/12/RC/2289 P2.

## References

Amgoud, L., and Ben-Naim, J. 2018. Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning* 99:39–55.

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation*. College Publications.

Baroni, P.; Rago, A.; and Toni, F. 2018. How many properties do we need for gradual argumentation? In *AAAI*, 1736–1743.

Bezou-Vrakatseli, E.; Cocarascu, O.; and Modgil, S. 2025. Can large language models understand argument schemes? In *ACL (Findings)*, 13666–13681.

Carstens, L., and Toni, F. 2015. Towards relation based argumentation mining. In *ArgMining*, 29–34.

Cayrol, C., and Lagasquie-Schiex, M. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *ECSQARU*, 378–389.

Cunningham, E.; Greene, D.; Cross, J.; and Rago, A. 2026. Evaluating LLM-driven summarisation of parliamentary debates with computational argumentation. *CoRR* abs/2604.19331.

Cunningham, E.; Cross, J.; and Greene, D. 2025. Identifying algorithmic and domain-specific bias in parliamentary debate summarisation. *CoRR* abs/2507.14221.

de Tarlé, L. D.; Bonzon, E.; and Maudet, N. 2022. Multiagent dynamics of gradual argumentation semantics. In *AAMAS*, 363–371.

Dore, D.; Faralli, S.; and Villata, S. 2025. Leveraging graph structural knowledge to improve argument relation prediction in political debates. In *ArgMining*, 74–86.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2):321–358.

Fabbri, A. R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; and Radev, D. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9:391–409.

Feger, M.; Boland, K.; and Dietze, S. 2025. Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments. In *ACL*, 23900–23915.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2025. Argumentative large language models for explainable and contestable claim verification. In *AAAI*, 14930–14939.

Furman, D. A.; Torres, P.; Rodríguez, J. A.; Letzen, D.; Martinez, M. V.; and Alemany, L. A. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *EMNLP (Findings)*, 2942–2956.

Gorur, D.; Rago, A.; and Toni, F. 2025. Can large language models perform relation-based argument mining? In *COLING*, 8518–8534.

Hautli-Janisz, A.; Kikteva, Z.; Siskou, W.; Górska, K.; Becker, R.; and Reed, C. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *LREC*, 3291–3300.

Hwang, H.; Gotlieb, M. R.; Nah, S.; and McLeod, D. M. 2007. Applying a cognitive-processing model to presidential debate effects: Postdebate news analysis and primed reflection. *Journal of Communication* 57(1):40–59.

Kryściński, W.; Keskar, N. S.; McCann, B.; Xiong, C.; and Socher, R. 2019. Neural text summarization: A critical evaluation. In *EMNLP-IJCNLP*, 540–551.

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

Li, H.; Wu, Y.; Schlegel, V.; Batista-Navarro, R.; Madusanka, T.; Zahid, I.; Zeng, J.; Wang, X.; He, X.; Li, Y.; and Nenadic, G. 2024. Which Side Are You On? A Multi-task Dataset for End-to-End Argument Summarisation and Evaluation. In *ACL (Findings)*, 133–150.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *KR*, 63–73.

Rago, A.; Li, H.; and Toni, F. 2023. Interactive explanations by conflict resolution via argumentative exchanges. In *KR*, 582–592.

Roush, A.; Shabazz, Y.; Balaji, A.; Zhang, P.; Mezza, S.; Zhang, M.; Basu, S.; Vishwanath, S.; and Shwartz-Ziv, R. 2024. OpenDebateEvidence: A massive-scale argument mining and summarization dataset. In *NeurIPS*, 34270–34293.

Ruiz-Dolz, R.; Gemechu, D.; Kikteva, Z.; and Reed, C. 2025. Looking at the unseen: Effective sampling of non-related propositions for argument mining. In *COLING*, 2131–2143.

Savigny, H., and Yun, B. 2026. AMELIA: A family of multi-task end-to-end language models for argumentation. *Argument Comput.* 17(1):79–103.

Tsfati, Y. 2003. Debating the debate: The impact of exposure to debate news coverage and its interaction with exposure to the actual debate. *Harvard International Journal of Press/Politics* 8(3):70–86.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating text generation with BERT. In *ICLR*.