

Have You Been Convinced? A Neuro-symbolic Argumentation Reasoning Approach to Detecting Persuasiveness in ChangeMyView Discussions

Siddharth Bhargava^{1,2}[0000–0002–2682–8557], Sara Tonelli¹[0000–0001–8010–6689],
and Patricia Martín-Rodilla³[0000–0002–1540–883X]

¹ Fondazione Bruno Kessler, Trento, Italy
{sbhargava,satonelli}@fbk.eu

² Universidade da Coruña, Coruña, Spain

³ IEGPS-CSIC, Spanish National Research Council Spain
p.m.rodilla@iegps.csic.es

Abstract. Argument structures represent how individual arguments interact within a discourse. However, leveraging these structures to infer meaningful knowledge remains challenging, as it requires a formal representation of how argumentative influence propagates across the structure. Moreover, large language models (LLMs) often produce opaque predictions, limiting trustworthiness in sensitive domains such as politics, law, and healthcare. We propose Neuro-symbolic Argumentation Reasoning System (NARS) for performing persuasiveness detection directly from online discussions. The neural component employs retrieval-augmented iterative LLM reasoning to extract bipolar argument structures from long discussions, while the symbolic component applies gradual argumentation semantics to propagate argumentative influence until convergence. Using the Webis-CMV-20 corpus, we construct CMV2BAS, a dataset of long, multi-threaded ChangeMyView discussions annotated with LLM-derived bipolar argument structures and labeled with delta events (view change vs. no change). We investigate whether gradual argumentation semantics can effectively detect persuasiveness, whether augmenting them with argument quality and graph-based centrality improves performance, and whether symbolic reasoning remains necessary in the presence of modern LLMs. NARS achieves F1=0.57 and AUC=0.56 under its best configuration, outperforming zero-shot Qwen3.6:27B and GPT-oss:20B, which fail to surpass random performance. We identify support-dominant discussions and long argumentative chains as key failure modes limiting attack influence propagation. Our work demonstrates the feasibility of operationalizing formal gradual argumentation semantics over noisy real-world discourse for scalable macro-level persuasion analysis, while highlighting key challenges and directions for future research in neuro-symbolic argumentation reasoning. We release both CMV2BAS dataset and the NARS framework for public use.⁴

Keywords: Neuro-symbolic reasoning · Argument mining · Persuasiveness detection · Social media.

⁴ Our open-source project repository: <https://github.com/The-obsrvr/NARS-Pipeline>

1 Introduction

Argument mining (AM), a core research area under Knowledge Representation and Reasoning, studies how arguments within a discourse are proposed, debated, and ultimately accepted. In dialogical settings, this becomes particularly important, as multiple participants take turns to present and challenge arguments with the goal of reaching a resolution or consensus [44, 29]. Analyzing argumentative interactions has applications in identifying key arguments, determining speaker stance, predicting discussion outcomes, and supporting explainable reasoning through interpretable decision traces [40, 23].

In real-world settings, AM relies on two important foundations: (i) *formal theoretical frameworks* that model argumentative interactions [15, 6], and (ii) *data-driven methods* that leverage discourse cues to identify and classify arguments [10, 23]. While substantial progress has been made in both formal argumentation [28] and computational argument mining [25, 11], effectively integrating these paradigms to produce interpretable, large-scale argument analysis remains an open problem [38].

Neuro-symbolic AI offers a promising direction for bridging formal and natural argumentation by combining the scalability of neural learning with the interpretability of symbolic reasoning [16]. Such approaches have demonstrated potential for improving reliability and explainability across knowledge-intensive tasks [42]. However, relatively little work has explored neuro-symbolic approaches for large-scale, real-world argument mining [38, 42], where systems must process raw, noisy discourse while producing formally grounded and traceable knowledge representations.

To address this gap, we propose a **Neuro-symbolic Argumentation Reasoning System** (NARS) consisting of two system components: a neural component that extracts bipolar argument structures from raw discussions using iterative LLM-based reasoning, and a symbolic component that applies gradual argumentation semantics to evaluate argumentative influence and derive structured knowledge. We evaluate NARS on the task of **persuasiveness detection**, where the objective is to determine whether the original user changes their viewpoint (henceforth referred to as the *delta event*) during a discussion. Experiments are conducted on our novel CMV2BAS Dataset, derived from the Webis-CMV-20 corpus [1], a large collection of moderated multi-threaded discussions from the ChangeMyView (CMV) subreddit.

To detect persuasiveness, our framework assigns each argumentative unit an *a priori* strength representing its initial persuasive influence, which is then iteratively propagated through gradual semantics until equilibrium is reached. Since these *a priori* strengths directly shape how influence evolves across the argument graph, we investigate multiple initialization strategies and analyze their impact on persuasiveness detection. Specifically, we address the following research questions:

RQ1. Can gradual argumentation semantics effectively model persuasiveness in long discussions?

RQ2. Does incorporating argument quality and structural centrality improve persuasion modeling under gradual semantics?

RQ3. Can large language models independently perform reliable persuasiveness reasoning, or are explicit symbolic reasoning frameworks necessary for modeling argumentative influence?

The main contributions of our work are:

1. **CMV2BAS Dataset:** A dataset of long, multi-threaded CMV discussions annotated with LLM-derived bipolar argument structures and delta events.
2. **NARS Framework:** An open-source, neuro-symbolic system for processing long discussions and determining persuasiveness.
3. **Analytical Study:** A systematic evaluation of *a priori strength* strategies and their impact on modeling argumentative influence and persuasiveness.

2 Background

2.1 Argumentation Semantics

Formal argumentation theory offers several families of semantics for evaluating arguments, differing in how they model acceptability. **Extension-based semantics** identifies sets of jointly acceptable arguments — known as extensions — satisfying properties such as admissibility, groundedness, or stability [13, 33]. While logic-driven, extensions yield binary accept/reject decisions and offer limited support for fine-grained argument comparison. **Ranking-based semantics** addresses this by inducing a total or partial ordering over arguments according to relative acceptability [2, 20], enabling richer comparisons even in the presence of multiple extensions. However, rankings do not quantify absolute argumentative strength and disregard intrinsic argument properties, limiting their applicability in real-world settings where arguments vary in quality and certainty.

Gradual semantics addresses these limitations by assigning continuous acceptability values to arguments based on structural interactions via a gradual iterative process [4, 8]. This makes gradual semantics well-suited for natural argumentative discourse, where arguments are inherently qualitative and uncertainty is pervasive, especially in dynamic, non-monotonic settings [45, 34]. Gradual semantics evaluates arguments through an iterative process that (i) aggregates supporting and attacking influences, and (ii) updates strength values until a stable equilibrium is reached [36, 7].

A critical yet underexplored factor in applying gradual semantics to real-world discourse is the *a priori* strength initialization of argument structures. Due to complexity of real-world discourse, prior work has been largely limited to controlled, synthetic settings with manually curated argument graphs [8, 36], leaving open the question of how to effectively initialize and propagate strength in large-scale argument structures. In this work, we directly address this gap: we propose and evaluate multiple initialization strategies grounded in both semantic quality and structural centrality, and apply gradual semantics at scale to real-world online discussions. This sets our framework NARS among the first concrete

steps toward operationalizing formal argumentation theory in practical, end-to-end argument analysis pipelines.

2.2 Persuasiveness Detection

Argumentation frameworks are well-suited for modeling non-monotonic reasoning, where accepted arguments may be revised as new evidence is introduced [5]. This makes them naturally applicable to tasks such as belief revision [14], polarity prediction [24], and persuasiveness detection [43, 18].

Persuasion is a central phenomenon in online discourse, particularly in domains such as politics, law, and debate [27]. Existing computational approaches typically model persuasiveness as a binary classification task using surface-level features or fine-tuned language models [41, 21, 22, 9]. However, persuasion in long-form discussions emerges gradually through multi-turn argumentative interactions rather than isolated discourse cues. Recent argumentation-aware approaches [38] explore persuasion at macro-level, though remain limited in their application and annotated datasets. To address this our proposed system, NARS, models persuasiveness detection as a gradual shift in the argument strength of the central proposition (the main opinion to change) through argumentative influence propagation, where the human participant interactions act as indicators of supporting and attacking signals. For example, multiple attacks against a claim may collectively reduce its strength even when individual arguments are only moderately persuasive. Similarly, a strong argument asserts a strong influence on to its target. In this work, we focus on operationalizing existing formal gradual argumentation semantics over noisy real-world discourse for scalable macro-level persuasion analysis

3 Data

We source our data from the publicly available Webis ChangeMyView Corpus 2020 (Webis-CMV-20) [1], which contains posts and comments from the ChangeMyView subreddit spanning 2005 to September 2017. The original corpus consists of 65,169 submissions and approximately 602,706 discussion threads replying directly to the original poster (OP). CMV is a heavily moderated subreddit with explicit discussion rules, where moderators actively remove any non-compliant comments.

For this study, we extract discussion threads satisfying several selection criteria to meet our experimental design. Each thread begins with an OP presenting a viewpoint intended to be challenged and consists of a tree-structured discussion containing replies and nested interactions. Threads are categorized into either **delta threads** (positive outcome), where the OP awards a explicit delta (Δ) indicating a change in its viewpoint, or **non-delta threads** (negative outcome), where no delta is awarded. Given non-delta threads do not have clear endpoint, unlike delta threads, long threads are manually ended on the comment containing the 500th sentence. Additionally to narrow our search space, we apply topic

filtering targeting discussions related to *health, migration, climate, and European integration*—topics that are politically polarizing and typically associated with higher argumentative activity. Finally, each thread must contain minimum five comments, spanning a minimum total of 30 sentences to ensure sufficient conversational activity.

All threads undergo preprocessing steps prior to analysis. Threads are pseudonymized to preserve user privacy, and no user profiling is performed in this study. Removed or deleted comments are retained to preserve reply structure integrity, although they are excluded from argument structure modeling. User mentions and quotations are replaced with the placeholder string “previously stated...” to maintain contextual relations between comments.

Following these steps, we identified 1,822 well-structured threads, of which 327 were delta threads. Due to the computational cost of our NARS pipeline, we sample 500 threads (250 delta- and 250 non-delta threads) for this study and release the broader data processing pipeline publicly.⁵ Delta threads contain an average of 11.39 turns (maximum 106) and 78.12 sentences (maximum 508), while non-delta threads contain an average of 13.33 turns (maximum 106) and 107.90 sentences (maximum 608). These threads are subsequently processed through the proposed NARS framework to form our CMV2BAS.

4 Methodology

Our proposed framework, NARS, consists of two primary components: a *neural* component and a *symbolic* component. The neural component employs a retrieval-augmented, iterative large language model (LLM) reasoning pipeline to extract coherent bipolar argument structures from multi-threaded discussions. These structures are subsequently processed by the symbolic component, which initializes each argumentative unit with an *a priori* strength value representing its initial influence. Iterative influence propagation is then performed over the resulting argument graph using gradual argumentation semantics until a stable equilibrium state is reached.

The resulting strength dynamics are used to model persuasiveness as a gradual argumentative process rather than a single binary event. Figure 1 illustrates the overall NARS architecture, highlighting the interaction between the neural and symbolic components as well as the downstream persuasiveness detection task. Overall, the framework is organized into **seven** modular stages spanning argument extraction, relation prediction, graph construction, symbolic reasoning, and persuasiveness detection (see Table 1 for the complete system implementation details).

Step 1. LLM-based EDU Extraction. *Elementary Discourse Units* are minimal discourse segments that express a single coherent proposition [39], typically split at clause boundaries and discourse connectives (e.g., *and, but, or*).

We use a reasoning-based LLM to segment each discussion into EDUs. This approach improves robustness over rule-based splitting by handling implicit or

⁵ <https://github.com/The-obsrvr/CMV-BAS-Data-Pipeline>

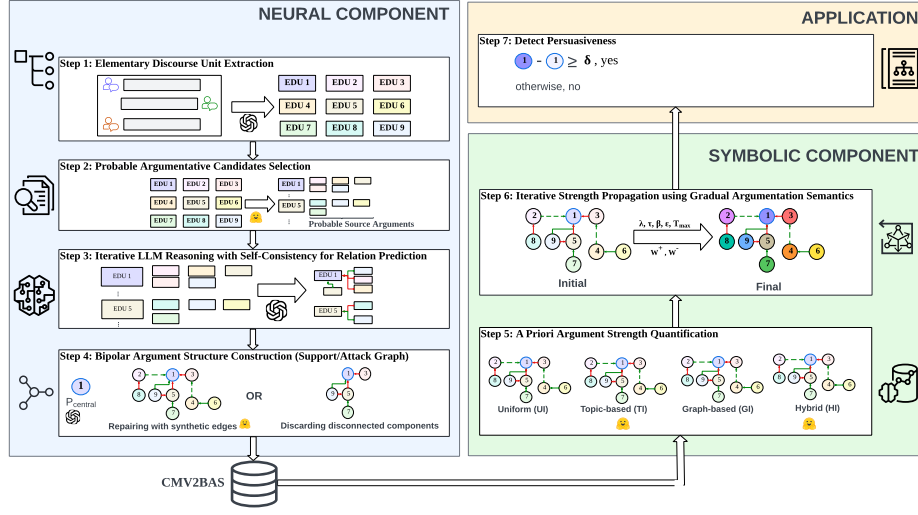


Fig. 1: Overview of the NARS architecture, illustrating the neural component and its steps, the symbolic component and the downstream persuasiveness detection module.

non-syntactic boundaries common in social media text. For example, “Climate change is real and it threatens our future” is split into [“Climate change is real”, “and it threatens our future”]. If the LLM generation fails after two tries, a punctuation-based fallback logic is performed ensures full coverage.

Step 2. Probable Argumentative Candidate Selection. Given the large size of individual discussion threads, open-source LLMs ($\leq 30B$) often exhibit reasoning errors and reduced task alignment when dealing with long contexts [26]. To mitigate this, we adopt a retrieval-augmented strategy in which, for each EDU (the target), we select argumentatively relevant source candidates, referred to as Probable Argumentative Candidates (PACs) before the LLM reasoning step.

PACs are retrieved by k -nearest-neighbour search over sentence-transformer embeddings [37], filtered by a minimum cosine similarity threshold, from which the top n_{semantic} candidates are retained. To capture contextually relevant but semantically dissimilar EDUs (e.g., immediate neighbours: [“...I believe that is why I don’t trust it”]), we additionally include up to n_{context} subsequent EDUs not already selected. Finally, temporal directionality is enforced by restricting source candidates to EDUs occurring after the target EDU.

Step 3. Iterative LLM Reasoning for Relation Prediction. Given PACs for each EDU, an LLM (in *thinking* mode) performs zero-shot classification to assign each candidate a *support*, *attack*, or *neutral* (non - argumentative) relation toward its target EDU. Additionally in the prompts, the context surrounding the target EDU, up to two sentences before and after, is included to facilitate LLM reasoning.

To improve reliability under noisy long-context inputs, we use self-consistency scheme with two low-temperature inference passes for deterministic generation, followed by majority voting. If any, the disagreements are resolved via a third tie-breaking call, providing a form of weak supervision without task-specific training [12]. We retain only EDUs, henceforth becoming argumentative units, that participate in at least one support or attack relation, filtering out non-argumentative or isolated units.

Step 4. Bipolar Argument Structure Construction. Predicted relations are consolidated into a bipolar argument graph $G(\mathcal{A}, \mathcal{E}, P_{\text{central}})$, where $\mathcal{A}$ denotes argumentative units, $\mathcal{E}$ support and attack relations, and P_{central} the original poster’s central proposition. P_{central} is identified via a medium-thinking LLM prompt over opening-post units and manually validated; we further ensure it carries at least one incoming edge to enable influence propagation.

Since Step 3 may yield disconnected components, we consider two consolidation strategies: *synthetic*, where disconnected components are linked to the main component via a single best similarity-based edge identified through cosine similarity over sentence-transformer embeddings, and *discarded*, where only the component containing P_{central} is retained. Graphs with fewer than three argumentative units are excluded, due to insufficient argumentative activity for meaningful evaluation.

Step 5. A Priori Argument Strength Quantification. Each argumentative unit in G carries some degree of persuasive influence within the discussion. To model how this influence affects the central proposition, arguments must first be assigned appropriate *a priori* strengths. Since arguments in online discussions are often implicit, incomplete, and vary in quality and structural importance, this initialization step critically captures these characteristics prior to influence propagation.

Formally, each argument graph is represented as an argumentation framework $AF = \langle \mathcal{A}, \mathcal{E}, \mathcal{I} \rangle$, where $\mathcal{A}$ denotes the set of arguments, $\mathcal{E}$ the set of support/attack relations, and $\mathcal{I}$ the initialization function assigning a priori argument strengths. We explore following complementary initialization strategies based on intrinsic and structural properties of the argument graph.

1. **Uniform Initialization (UI):** Following standard gradual semantics [34], each argumentative unit is assigned an identical prior strength of 1.0 (see Algorithm 1, lines 1–4), independent of its content or structure. This provides a neutral baseline where no prior bias is introduced on argumentative influence and serves as the baseline for **RQ1**.
2. **Topic-based Initialization (TI):** A priori strengths are estimated using a DeBERTa-v3-small [19] pairwise argument-quality model fine-tuned on the Webis-ArgQuality-20 corpus [17], which contains human-annotated relevance scores. This formulation explicitly scores each argumentative unit with respect to P_{central} , using topical relevance as a substitute for persuasive influence. While effective for explicit, well-formed arguments aligned with P_{central} , it is less reliable for implicit or context-dependent units (see Algorithm 1, lines 5–9).

Algorithm 1 A Priori Argument Strength Quantification

Require: Units $\mathcal{A} = \{a_1, \dots, a_n\}$; relations $\mathcal{E}$; fine-tuned argument quality model; central proposition P_{central}

Ensure: Normalized node strengths $S = \{s_1, \dots, s_n\}$

- 1: **UI scoring** (u'_i):
- 2: **for** each unit a_i **do**
- 3: Set $u'_i = 1$
- 4: **end for**
- 5: **TI scoring** (r'_i):
- 6: **for** each unit a_i **do**
- 7: Compute pairwise quality score r_i using fine-tuned argument quality model with input $(P_{\text{central}}, a_i)$
- 8: **end for**
- 9: Normalize r_i to obtain $r'_i \in [0, 1]$
- 10: **GI scoring** (c'_i):
- 11: Construct directed graph $G = (\mathcal{A}, \mathcal{E})$
- 12: Compute PageRank centrality c_i for each unit a_i
- 13: Normalize c_i to obtain $c'_i \in [0, 1]$
- 14: **HI scoring** (h'_i):
- 15: Compute undirected shortest path distance d_i from each a_i to P_{central} in G
- 16: Set proximity weight $\alpha(d_i) = 1 - d_i/d_{\text{max}}$
- 17: $h_i = r'_i + \alpha(d_i) \cdot c'_i$
- 18: Normalize h_i to obtain $h'_i \in [0, 1]$
- 19: **return** $S = \{(u'_i, r'_i, c'_i, h'_i)\}$

3. **Graph-based Initialization (GI):** A priori strengths are computed using graph-based centrality measures, specifically PageRank in-degree centrality [32], to capture the structural importance of the units within the argument graph. This strategy emphasizes arguments occupying influential structural positions irrespective of semantic quality, making it effective for identifying structurally central arguments but potentially weaker in semantic interpretation (see Algorithm 1, lines 10–13).
4. **Hybrid Initialization (HI):** Motivated by the idea that argumentative importance depends on both semantic relevance and graph structure, HI combines topic-based quality (TI) and structural relevance (GI) with proximity to P_{central} . It preserves semantic strength while incorporating a distance-weighted structural bias, prioritizing arguments that are both semantically relevant and centrally positioned in the graph as the most influential. Formally,

$$\text{HI}(a) = \text{normalise}(\text{TI}(a) + \alpha(d) \cdot \text{GI}(a)) \quad (1)$$

where $\alpha(d) = 1 - \frac{d}{d_{\text{max}}}$ and d is the shortest path distance to P_{central} . This ensures that structural influence is strongest near P_{central} and diminishes for peripheral arguments, which typically have weaker structural impact (see Algorithm 1, lines 14–18).

Algorithm 2 Iterative Strength Propagation using Gradual Semantics

Require: Nodes $\mathcal{A} = \{a_1, \dots, a_n\}$; support weights $W^+ \in \mathbb{R}_{\geq 0}^{n \times n}$; attack weights $W^- \in \mathbb{R}_{\geq 0}^{n \times n}$; prior strengths $b \in [0, 1]^n$; damping $\lambda \in (0, 1]$; temperature $\tau > 0$; beta $\beta \geq 0$; tolerance $\varepsilon > 0$; maximum iterations $T_{\max}$

Ensure: Strength scores $s^* \in [0, 1]^n$; trajectory $\{s^{(t)}\}$

- 1: **Initialization:** $s^{(0)} \leftarrow b$; $t \leftarrow 0$
- 2: **repeat**
- 3: Compute aggregated support and attack:

$$p^{(t)} = (W^+)^{\top} s^{(t)}, \quad q^{(t)} = (W^-)^{\top} s^{(t)}$$
- 4: Net influence: $u^{(t)} = \beta b + p^{(t)} - q^{(t)}$
- 5: Gradual function (sigmoid): $\sigma_{\tau}(x) = \frac{1}{1 + e^{-x/\tau}}$
- 6: Proposed update: $\tilde{s}^{(t+1)} = \sigma_{\tau}(u^{(t)})$
- 7: Damped iterate: $s^{(t+1)} = (1 - \lambda) s^{(t)} + \lambda \tilde{s}^{(t+1)}$
- 8: $t \leftarrow t + 1$
- 9: **until** $\|s^{(t)} - s^{(t-1)}\|_{\infty} < \varepsilon$ **or** $t \geq T_{\max}$
- 10: **Return** $s^* \leftarrow s^{(t)}$

Additionally, we consider a **Majority Vote (Ensemble)** strategy, which aggregates the binary persuasiveness predictions across all four initialization strategies via majority vote, predicting persuasiveness when at least two strategies vote yes.

Besides node initialization, relations are also initialized through weighted support and attack matrices, denoted by $W^+ \in \mathbb{R}_{\geq 0}^{n \times n}$ and $W^- \in \mathbb{R}_{\geq 0}^{n \times n}$ respectively, encoding the degree to which arguments positively or negatively influence one another. Due to the substantial imbalance between support and attack relations in the extracted argument structures, attacking arguments may be overwhelmed by the support majority during propagation. To counter this, we initialize the attack matrix W^- with a $3 \times$ boosting factor, amplifying their influence in the propagation.

Step 6. Iterative Strength Propagation using Gradual Semantics.

We apply gradual semantics [3, 7] (Algorithm 2) to compute posterior (final) argument strengths. At each iteration, the framework *aggregates* the supporting and attacking influences via the relation weights, *propagates* them across the graph, and *updates* argument strengths through a sigmoid-based gradual function [34]. This repeats until strength values converge to a stable equilibrium, enabling a macro-level, context-sensitive evaluation of persuasive influence across the full argument structure.

The algorithm is governed by five hyperparameters: λ controls how much newly propagated information updates current strengths (line 6); β retains a fraction of the a priori strength throughout propagation (line 4); τ governs the sharpness of the gradual update function (line 5); and ε , $T_{\max}$ define the convergence criterion (line 7). Values have been taken from existing semantics and

tuned empirically (see Table 1) and we leave a more detailed investigation of their effects for future work.

Step 7. Persuasiveness Detection. Persuasiveness is modeled as a *significant posterior reduction in the graded strength value of the central proposition within the given argumentation framework*. Let b_{central} and s_{central}^* denote the prior and posterior strengths of P_{central} after the iterative symbolic reasoning. A discussion is considered persuaded if the change in strength exceeds a predefined **persuasiveness sensitivity** threshold, denoted by δ :

$$\text{Persuaded}, |\Delta(P_{\text{central}})| = \begin{cases} \text{yes}, & \text{if } |s_{\text{central}}^* - b_{\text{central}}| \geq \delta \\ \text{no}, & \text{otherwise} \end{cases}$$

Larger values of δ (≥ 0.5) impose stricter, more conservative criteria for persuasiveness, whereas smaller values may allow even subtle shifts in P_{central} ’s argumentative strength to be considered sufficiently persuasive. In this work, we analyze the effect of varying δ on persuasiveness detection.

5 Experimental Setup

This study primarily focuses on Step 5 of our framework, where we investigate how *a priori* argument strength can be modeled to capture the persuasive influence of argumentative units and subsequently propagated through the argument structure to infer persuasiveness. We implement the four initialization strategies introduced in Step 5 of Section 4 and evaluate their effectiveness in addressing **RQ1** and **RQ2**.

To address **RQ3**, we further investigate whether modern LLMs can independently perform persuasiveness detection without explicit symbolic reasoning mechanisms. Specifically, we experiment with two state-of-the-art, open-source thinking LLMs ($\leq 30B$, due to computational limit), *Qwen3.6:27B* [35] and *GPT-oss:20B* [31] (medium thinking), under three zero-shot prompting settings: (i) using only the discussion text, (ii) using the discussion together with the bipolar argument structure (synthetic version) generated in Step 4, and (iii) additionally incorporating the *a priori* argumentative strength values (HI) produced in Step 5. Although prompts are adapted to each setting, the models are constrained to output only a binary persuasiveness decision.

Given the lack of existing work on this data and task, we benchmark our framework against two naive baselines: a random classifier and a feature-based logistic regression model using surface-level discussion properties. The logistic regression baseline uses four features—direct attacks on P_{central} , total attacks, support–attack ratio, and the in-degree of P_{central} —to estimate persuasiveness. To train this baseline, we additionally processed 100 discussions from the original corpus.

All experiments are conducted on a 48GB NVIDIA Ampere A40 GPU within a Docker-based environment. For the steps involving LLMs, the models and their inference are performed through the Ollama client [30]. Table 1 summarizes

Step	Configuration
S1 EDU Extraction	Qwen3:27B [35]; one-shot, zero-thinking; temp=0.1; batch=200 sentences; max 8,192 tokens; punctuation fallback
S2 PAC Selection	all-MiniLM-L6-v2 [37]; $k=25$; sim. threshold=0.45; $n_{\text{semantic}}=10$, $n_{\text{context}}=5$
S3 Relation Prediction	GPT-oss:20B (medium thinking) [31]; batch=10 EDUs ± 2 sentences for context; self-consistency with temp={0.1, 0.2, 0.15 (tie-breaker)}; max 8,192 tokens; JSON structured output
S4 BAS Construction	P_{central} via GPT-oss:20B (medium thinking); <i>discarded</i> : only main component containing P_{central} ; <i>synthetic</i> : cosine-similarity repair via all-MiniLM-L6-v2; min. 3 nodes in final graph
S5 Strength Init.	UI / GI / TI / HI / Ensemble; $W^+=1.0/W^-=3.0$; synthetic edge weight=0.7
S6 Gradual Semantics	$\lambda=0.5$, $\tau=0.2$, $\beta=0.1$, $\varepsilon=1\text{e-}6$, $T_{\text{max}}=1,000$
S7 Persuasiveness	$\delta \in \{0.1, 0.3, 0.5, 0.7, 0.8\}$

Table 1: Implementation details per pipeline step, reflecting the most efficient configuration identified through empirical investigation.

our final, most-efficient system configuration for each step. Full implementation details and source code are available in our system repository (see Footnote 4).

5.1 CMV2BAS Dataset

We construct and release CMV2BAS as a large-scale benchmark for persuasive argumentation analysis and neuro-symbolic reasoning. Starting from 500 raw discussion threads, we first segmented the text into 45,689 Elementary Discourse Units (EDUs), enabling fine-grained modeling of long and noisy social media discourse. Step 2 retrieved 350,101 probable argumentative candidate pairs (PACs) based on the semantic and contextual similarity. Step 3 performs iterative LLM-based relation prediction with self-consistent reasoning to classify support and attack relations, forming the computational bottleneck (9–12 minutes per 10 EDUs). Step 4 consolidates these predictions into connected bipolar argument structures centered on the central proposition, filtering out the smaller (≤ 3 units) discussion threads.

This pipeline yields 384 valid discussions from 500 CMV threads (202 delta-event, 182 non-delta-event). The final dataset contains 24,407 argumentative units and 48,166 support/attack relations. Both synthetic and discarded graph variants are retained in the dataset. Each discussion is enriched with metadata including the central proposition, speaker participation statistics, conversational

structure, symbolic reasoning outputs, and persuasiveness labels. Outputs were sampled after each step for manual validation and found to be largely stable, with minor inconsistencies handled in post-processing.

In summary, CMV2BAS is a large-scale benchmark of long-form persuasive discussions with structured argument graphs and symbolic reasoning traces, supporting evaluation of argument structure modeling and persuasiveness detection in complex discourse.

6 Results

We evaluate different modeling paradigms on persuasiveness detection, formulated as a binary classification task, using Accuracy, F1-score, and Area Under the ROC Curve (AUC). Accuracy measures the overall correctness, F1-score captures the precision-recall trade-off, and AUC evaluates the model’s discriminative power between persuasive and non-persuasive discussions. Since the symbolic propagation mechanism is deterministic given a fixed argument graph and initialization strategy, the framework does not exhibit variance across random seeds. Instead, to assess robustness under dataset variability, we report bootstrap confidence intervals ($B=1000$) [46] over evaluation metrics.

Beyond classification performance, we analyze the effect of the persuasiveness sensitivity δ and study how NARS behaves under varying δ . We additionally perform qualitative analysis on the generated argument structures and symbolic reasoning traces to identify common failure modes, including dominant support relations and signal attenuation across long argumentative paths.

6.1 Persuasiveness Detection

Table 2 reports the performance of different modeling paradigms on persuasiveness detection.

RQ1 — Gradual argumentation semantics for persuasiveness. NARS with Uniform Initialization (UI) shows that gradual argumentation semantics can capture persuasiveness signals beyond naive baselines. At $\delta = 0.3$, UI achieves $F1=0.5196$ [0.455, 0.580] and $AUC=0.5643$ [0.503, 0.618] under the synthetic variant, with the AUC interval lying above random performance (0.500). The wide confidence intervals indicate a modest signal, and results are broadly comparable to the logistic regression baseline, suggesting that standard gradual semantics captures a weak but non-trivial persuasiveness signal.

RQ2 — Role of argument quality and structural centrality. Initialization strategies have a limited and inconsistent effect on performance. UI achieves the best AUC despite its simplicity, and overlapping confidence intervals across all strategies strong conclusions. GI yields a higher F1 than UI (0.5493 vs. 0.5196), while TI and HI do not consistently improve over simpler alternatives. The Ensemble achieves the highest F1 overall (0.5682), though its interval overlaps with other strategies. Synthetic graph variants show marginal gains

Category	Method	Accuracy	F1	AUC
Naive Baseline	Random	0.5025	0.5266	0.5000
	Log. Regression	0.5392	0.5635	0.5869
NARS (discarded); $\delta = 0.3$	UI	0.5417 ^{0.591} _{0.489}	0.5165 ^{0.575} _{0.451}	0.5620 ^{0.616} _{0.502}
	GI	0.5000 ^{0.549} _{0.451}	0.5450 ^{0.604} _{0.486}	0.4697 ^{0.529} _{0.413}
	TI	0.4635 ^{0.513} _{0.411}	0.4976 ^{0.556} _{0.434}	0.4638 ^{0.525} _{0.409}
	HI	0.4505 ^{0.505} _{0.401}	0.4548 ^{0.517} _{0.392}	0.4605 ^{0.516} _{0.404}
	Ensemble	0.4922 ^{0.542} _{0.443}	0.5412 ^{0.601} _{0.482}	0.4757 ^{0.533} _{0.421}
NARS (synthetic); $\delta = 0.3$	UI	0.5521 ^{0.602} _{0.502}	0.5196 ^{0.580} _{0.455}	0.5643 ^{0.618} _{0.503}
	GI	0.5000 ^{0.549} _{0.453}	0.5493 ^{0.608} _{0.491}	0.4691 ^{0.530} _{0.411}
	TI	0.4609 ^{0.513} _{0.411}	0.5106 ^{0.568} _{0.446}	0.4701 ^{0.526} _{0.411}
	HI	0.4896 ^{0.539} _{0.440}	0.5051 ^{0.562} _{0.440}	0.5108 ^{0.566} _{0.454}
	Ensemble	0.4974 ^{0.549} _{0.448}	0.5682 ^{0.621} _{0.510}	0.5135 ^{0.573} _{0.458}
Qwen3.6:27B (0-shot)	Discussion	0.4914	0.2525	0.5227
	+ structure (synth)	0.4845	0.2647	0.5140
	+ HI prior strength	0.4880	0.2660	0.5178
GPT-oss:20B (0-shot)	Discussion	0.4674	0.1341	0.5041
	+ structure (synth)	0.4742	0.1639	0.5099
	+ HI prior strength	0.4708	0.1250	0.5085

Table 2: Evaluation of persuasiveness detection across modeling paradigms on CMV2BAS ($n = 384$). NARS results report mean with 95% confidence interval through bootstrapping (^{upper}_{lower}, $B=1000$) at $\delta = 0.3$. LLM baselines report point estimates only. **Bold** denotes best per column among NARS configurations.

over discarded variants under UI and GI, being more structurally oriented, with negligible differences in other configurations.

RQ3 — LLM-based vs. symbolic reasoning. Both LLMs, with thinking mode, perform well below the random baseline in F1 (Qwen: 0.2660, GPT-oss: 0.1639 at best), with AUC near 0.50. Augmenting prompts with argument structure and strength scores yields only marginal gains, suggesting that zero-shot LLMs struggle to exploit structured argumentation representations for this task.

The large bootstrap confidence intervals (typically ± 0.05 – 0.06) reflect the difficulty of the task and the moderate corpus size. Overlapping intervals across NARS configurations indicate that differences between individual strategies are not statistically pronounced, with structural features appearing to be the dominant factor for the initialization choice.

6.2 Persuasiveness Sensitivity

To evaluate the robustness of NARS under varying persuasiveness sensitivity thresholds, we analyse the relationship between δ , F1-score, and the predicted positive rate – the rate at which the system classifies a discussion as persuasive. Figure 2 presents the results.

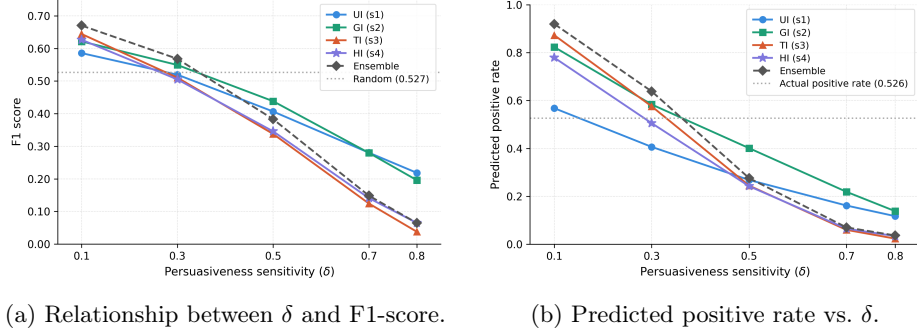


Fig. 2: Effect of persuasiveness sensitivity threshold δ on F1 score (left) and prediction bias (right) across initialization strategies in synthetic mode ($n = 384$).

At lower values of δ , all initialization strategies exhibit a substantially inflated tendency to predict the positive class. As shown in Figure 2b, most strategies intersect the true positive rate (0.526) between $\delta = 0.3$ and $\delta = 0.5$, indicating that this range constitutes the most calibrated operating region for persuasiveness detection. As δ increases, the framework becomes progressively conservative, requiring stronger accumulated argumentative influence to classify a discussion as persuasive, and F1 degrades accordingly as the system defaults toward the negative class.

Among the initialization strategies, UI followed by GI demonstrate the most stable F1 and AUC across the sensitivity range. In contrast, TI, HI, and the Ensemble exhibit a sharp performance decline beyond $\delta = 0.5$, suggesting that semantic and hybrid initializations propagate insufficient influence to maintain discriminability at stricter thresholds. Overall, δ acts as a tunable hyperparameter governing the precision–recall trade-off, with $\delta = 0.3$ offering the most balanced configuration among our experiments.

Error Analysis. Through qualitative analysis, we identify two primary failure modes that limit effective attack signal propagation. *Support dominance:* in highly support-dominant discussions, the overwhelming volume of supporting arguments suppresses any influence of attacking arguments on P_{central} , regardless of their argumentative quality. To partially mitigate this effect, we experimentally applied a $3\times$ weight multiplier to attack edges; however, the resulting performance gains remained marginal, suggesting that the opposing signals were insufficient against the large volume of supporting signals to meaningfully influence the final propagation dynamics.

Signal attenuation: in long argumentative chains, the influence of peripheral attack nodes progressively weakens before reaching P_{central} , reducing their contribution to the final acceptability score. Both failure modes highlight a broader open challenge in modeling argumentative attacks within real-world discourse.

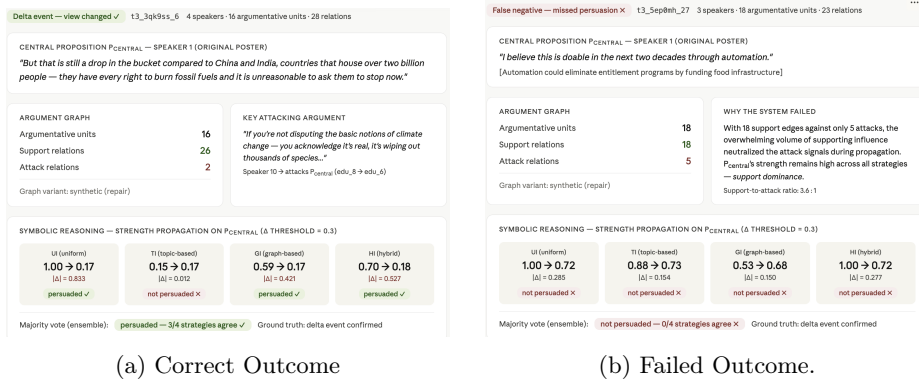


Fig. 3: NARS analysis on two discussion threads

7 Conclusion

We presented NARS, a neuro-symbolic system for argument structure extraction and macro-level persuasiveness detection over long, multi-threaded online discussions. We introduced CMV2BAS, a dataset of 384 discussions annotated with LLM-derived bipolar argument structures and delta events, which we release publicly alongside the framework. Figure 3 presents an infographic illustrating the NARS analysis of discussion threads taken from CMV2BAS. In summary, NARS achieves $F1=0.57$ and $AUC=0.56$ at $\delta = 0.3$, moderately outperforming both feature-based and random baselines, while zero-shot LLMs fail to surpass random performance even when augmented with extracted argument structures and strength scores. Although NARS remains limited in practical deployment, our results demonstrate the feasibility of applying formal argumentation semantics to noisy real-world discourse for scalable macro-level persuasion analysis.

Our study underscores that strength initialization remains the primary bottleneck of the neuro-symbolic framework. Quantifying the influence and relative strength of arguments is a challenging problem, particularly when structural and quality-based initialization may introduce noisy or conflicting propagation patterns. In addition, the current NARS framework exhibits performance degradation over large argumentative distances, leading to signal attenuation across long argument chains.

Future work should explore recursive and hierarchical reasoning and more structure-oriented initialization strategies in which smaller argumentative substructures are processed independently before aggregation, helping preserve local argumentative activity and influence signals in increasingly large and complex discussions.

Acknowledgments. This research work has received funding from the European Union’s Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351.

References

1. Al-Khatib, K., Völske, M., Syed, S., Kolyada, N., Stein, B.: Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In: 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). pp. 7067–7072. Association for Computational Linguistics (Jul 2020). <https://aclanthology.org/2020.acl-main.632/>
2. Amgoud, L., Ben-Naim, J.: Ranking-based semantics for argumentation frameworks. In: Liu, W., Subrahmanian, V.S., Wijsen, J. (eds.) *Scalable Uncertainty Management*. pp. 134–147. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
3. Amgoud, L., Ben-Naim, J.: Weighted Bipolar Argumentation Graphs: Axioms and Semantics. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. pp. 5194–5198. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden (Jul 2018). <https://doi.org/10.24963/ijcai.2018/720>
4. Amgoud, L., Doder, D., Lagasquie-Schiex, M.C.: Higher-Order Argumentation Frameworks: Principles and Gradual Semantics. In: *Thirty-Third International Joint Conference on Artificial Intelligence*. vol. 4, pp. 3224–3231 (Aug 2024). <https://doi.org/10.24963/ijcai.2024/357>
5. Amgoud, L., Nouioua, F.: An argumentation system for defeasible reasoning. *International Journal of Approximate Reasoning* **85**, 1–20 (Jun 2017). <https://doi.org/10.1016/j.ijar.2017.02.009>
6. Baroni, P., Gabbay, D.M., Giacomin, M., vander Torre, L. (eds.): *Handbook of Formal Argumentation*. College Publications (2018)
7. Baroni, P., Rago, A., Toni, F.: From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning* **105**, 252–286 (Feb 2019). <https://doi.org/10.1016/j.ijar.2018.11.019>
8. Beuselinck, V., Delobelle, J., Vesic, S.: A Principle-based Account of Self-attacking Arguments in Gradual Semantics. *Journal of Logic and Computation* **33**(2), 230–256 (Mar 2023). <https://doi.org/10.1093/logcom/exac093>
9. Bozdag, N.B., Mehri, S., Yang, X., Ha, H., Cheng, Z., Durmus, E., You, J., Ji, H., Tur, G., Hakkani-Tür, D.: Must Read: A Systematic Survey of Computational Persuasion (May 2025). <https://doi.org/10.48550/arXiv.2505.07775>
10. Budzynska, K., Janier, M., Reed, C., Saint-Dizier, P.: Theoretical foundations for illocutionary structure parsing. *Argument & Computation* **7**(1), 91–108 (Jan 2016). <https://doi.org/10.3233/AAC-160005>
11. Cabrio, E., Villata, S.: Five Years of Argument Mining: A Data-driven Analysis. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. pp. 5427–5433. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden (Jul 2018). <https://doi.org/10.24963/ijcai.2018/766>
12. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: *Proceedings of the 41st International Conference on Machine Learning. ICML’24*, vol. 235, pp. 11733–11763. JMLR.org, Vienna, Austria (Jul 2024)
13. Dung, P.M.: On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n -person games. *Artificial Intelligence* **77**(2), 321–357 (Sep 1995). [https://doi.org/10.1016/0004-3702\(94\)00041-X](https://doi.org/10.1016/0004-3702(94)00041-X)

14. Falappa, M.A., García, A.J., Kern-Isberner, G., Simari, G.R., Falappa, M.A., García, A.J., Kern-Isberner, G., Simari, G.R.: On the evolving relation between Belief Revision and Argumentation. *The Knowledge Engineering Review* **26**(1), 35–43 (Feb 2011). <https://doi.org/10.1017/S0269888910000391>
15. Gabbay, D., Giacomini, M., Guillermo, S.R., Thimm, M. (eds.): *Handbook of Formal Argumentation*, Volume 2, vol. 2 (Aug 2021)
16. d’Avila Garcez, A., Lamb, L.C.: Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review* **56**(11), 12387–12406 (Nov 2023). <https://doi.org/10.1007/s10462-023-10448-w>
17. Gienapp, L., Stein, B., Hagen, M., Potthast, M.: Efficient Pairwise Annotation of Argument Quality. In: *The 58th annual meeting of the Association for Computational Linguistics (ACL)*. ACL (Jul 2020)
18. Habernal, I., Gurevych, I.: Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 1589–1599. Association for Computational Linguistics, Berlin, Germany (2016). <https://doi.org/10.18653/v1/P16-1150>
19. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing (2021)
20. Heyninck, J., Raddaoui, B., Straßer, C.: Ranking-based Argumentation Semantics Applied to Logical Argumentation. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. pp. 3268–3276. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China (Aug 2023). <https://doi.org/10.24963/ijcai.2023/364>
21. Hidey, C., Musi, E., Hwang, A., Muresan, S., McKeown, K.: Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In: *Proceedings of the 4th Workshop on Argument Mining*. pp. 11–21. Association for Computational Linguistics, Copenhagen, Denmark (2017). <https://doi.org/10.18653/v1/W17-5102>
22. Ji, L., Wei, Z., Hu, X., Liu, Y., Zhang, Q., Huang, X.: Incorporating Argument-Level Interactions for Persuasion Comments Evaluation using Co-attention Model. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) *Proceedings of the 27th International Conference on Computational Linguistics*. pp. 3703–3714. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018)
23. Lawrence, J., Reed, C.: Argument Mining: A Survey. *Computational Linguistics* **45**(4), 765–818 (Jan 2020). https://doi.org/10.1162/coli_a_00364
24. Lenz, M., Bergmann, R.: PolArg: Unsupervised Polarity Prediction of Arguments in Real-Time Online Conversations. In: Cimiano, P., Frank, A., Kohlhase, M., Stein, B. (eds.) *Robust Argumentation Machines*. pp. 108–126. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-63536-6_7
25. Lippi, M., Torroni, P.: Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology* **16**(2), 1–25 (Apr 2016). <https://doi.org/10.1145/2850417>
26. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* **12**, 157–173 (2024). https://doi.org/10.1162/tacl_a_00638
27. Lukin, S., Anand, P., Walker, M., Whittaker, S.: Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In: Lapata, M., Blunsom, P., Koller, A. (eds.) *Proceedings of the 15th Conference of the European Chapter of*

- the Association for Computational Linguistics: Volume 1, Long Papers. pp. 742–753. Association for Computational Linguistics, Valencia, Spain (Apr 2017)
28. Modgil, S., Prakken, H.: The ASPIC+ framework for structured argumentation: A tutorial. *Argument & Computation* **5**(1), 31–62 (Jan 2014). <https://doi.org/10.1080/19462166.2013.869766>
 29. Morio, G., Fujita, K.: Annotating Online Civic Discussion Threads for Argument Mining. In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). pp. 546–553 (Dec 2018). <https://doi.org/10.1109/WI.2018.00-39>
 30. Ollama: Ollama. <https://ollama.com> (2025)
 31. OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R.K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K.G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E.Y., Leclerc, G., Lennon, J.P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimplouras, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., Zhao, S.: gpt-oss-120b & gpt-oss-20b model card (2025), <https://arxiv.org/abs/2508.10925>
 32. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking : Bringing Order to the Web. In: The Web Conference (Nov 1999)
 33. Polberg, S.: Extension-based semantics of abstract dialectical frameworks. In: STAIRS 2014, pp. 240–249. IOS Press (2014)
 34. Potyka, N.: Interpreting Neural Networks as Quantitative Argumentation Frameworks. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(7), 6463–6470 (May 2021). <https://doi.org/10.1609/aaai.v35i7.16801>
 35. Qwen Team: Qwen3.6-27B: Flagship-level coding in a 27b dense model (April 2026), <https://qwen.ai/blog?id=qwen3.6-27b>
 36. Rago, A., Vasileiou, S.L., Toni, F., Son, T.C., Yeoh, W.: A Methodology for Incompleteness-Tolerant and Modular Gradual Semantics for Argumentative Statement Graphs (Aug 2025). <https://doi.org/10.48550/arXiv.2410.22209>
 37. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (Nov 2019)
 38. Ruiz-Dolz, R., Heras, S., García-Fornes, A.: Automatic Debate Evaluation with Argumentation Semantics and Natural Language Argument Graph Networks (2022). <https://doi.org/10.48550/ARXIV.2203.14647>
 39. Saveleva, E., Petukhova, V., Mosbach, M., Klakow, D.: Discourse-based Argument Segmentation and Annotation. In: *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. pp. 41–53 (2021)

40. Shi, L., Giunchiglia, F., Wang, H., Cheng, Y., Song, R., Shi, D., Diao, X., Xu, H.: From text mining to intelligent debate: Task frameworks and technological evolution in computational argumentation. *Information Processing & Management* **63**(2, Part B), 104465 (Mar 2026). <https://doi.org/10.1016/j.ipm.2025.104465>
41. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., Lee, L.: Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In: *Proceedings of the 25th International Conference on World Wide Web*. pp. 613–624. WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (Apr 2016). <https://doi.org/10.1145/2872427.2883081>
42. Ulbricht, M., Wallner, J.: Strongly accepting subframeworks: Connecting abstract and structured argumentation. *Frontiers in Artificial Intelligence and Applications* **353**, 320–331 (2022). <https://doi.org/10.3233/FAIA220163>
43. Verma, B., Marreddy, M., Mehler, A.: Predicting Convincingness in Political Speech: How Emotional Tone Shapes Persuasive Strength. In: Barnes, J., Barriere, V., De Clercq, O., Klinger, R., Nouri, C., Nozza, D., Singh, P. (eds.) *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*. pp. 37–51. Association for Computational Linguistics, Rabat, Morocco (Mar 2026)
44. Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., Reed, C.: Argumentation in the 2016 US presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation* **54**(1), 123–154 (Mar 2020). <https://doi.org/10.1007/s10579-019-09446-8>
45. Wang, Z., Shen, Y.: Bilateral Gradual Semantics for Weighted Argumentation. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**(9), 10732–10739 (Mar 2024). <https://doi.org/10.1609/aaai.v38i9.28945>
46. Wood, M.: Statistical inference using bootstrap confidence intervals. *Significance* **1**(4), 180–182 (2004). <https://doi.org/https://doi.org/10.1111/j.1740-9713.2004.00067.x>, <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2004.00067.x>