Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning
KR meets Machine Learning and Explanation

Learnable Multi-Attribute Gradual Semantics for Predicting Persuasion in Argumentative Debates

Nino Pireaud[1] , **Victor David**[1] , **Anthony Hunter**[2] , **Pierre Monnin**[1] , **Elena Cabrio**[1]
[1]Université Côte d'Azur, Inria, CNRS, France
[2]University College London, UK
{nino.pireaud,victor.david,pierre.monnin,elena.cabrio}@inria.fr, anthony.hunter@ucl.ac.uk

Abstract

Gradual semantics for weighted bipolar argumentation provide a principled framework for modelling argumentative reasoning, yet existing approaches remain mostly scalar, fixed, and weakly grounded in empirical data. We introduce *learnable multi-attribute gradual semantics* for persuasion prediction in argumentative debates. Our approach builds a dataset of 600 textual debates converted into multi-attribute argumentation graphs enriched with multi-dimensional features on nodes and relations. Building on this representation, we propose learnable aggregation operators that distinguish intrinsic quality from persuasive strategy dimensions. Experiments show that the learned semantics achieve competitive performance with neural and LLM-based baselines while preserving interpretability.

1 Introduction

Debating is a fundamental human activity for exploring controversial issues and confronting opposing viewpoints. Beyond analysing individual arguments, a key challenge is to understand whether a debate leads participants to revise their stance, and if so, in which direction and to what extent. Ideally, predicting persuasion requires modelling it as a multi-dimensional phenomenon, capturing argumentative quality, rhetorical and persuasive strategies, and the structure of argumentative interactions through support and attack relations.

Recent work in NLP has made substantial progress on modelling persuasion and attitude change from dialogue data. Early studies identified lexical and interaction patterns associated with convincing arguments (Tan et al. 2016), proposed models to rank arguments by perceived convincingness (Habernal and Gurevych 2016), and analysed the semantic roles of claims and premises in online debates (Hidey et al. 2017). More recent research has explored higher-level discourse properties, showing for instance that storytelling strategies and narrative framing correlate with persuasion outcomes in ChangeMyView discussions (Nabhani et al. 2025), and that the ordering and coherence of arguments play a crucial role in persuasion success (Mirzakhmedova et al. 2023). Together, these works demonstrate that stance evolution can often be inferred from textual and conversational features.

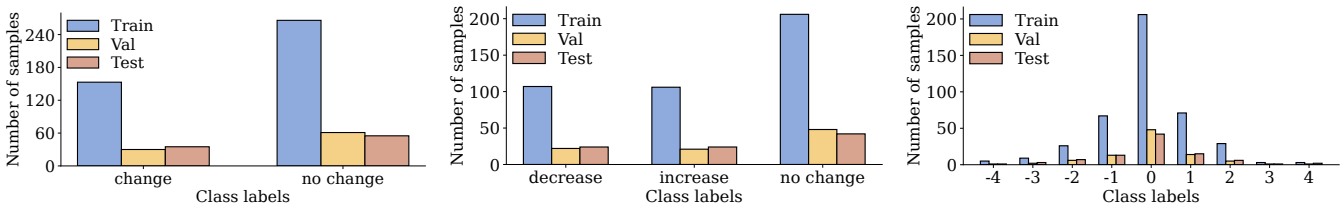
However, most existing NLP approaches model persuasion as a direct mapping from dialogue representations to outcomes, typically relying on sequential encoders or bag-of-features representations. While effective, such models do not explicitly capture the underlying *argumentative structure* of debates, namely how claims support or attack each other and how multiple qualitative factors (e.g., evidentiality, emotional tone, or confidence) interact during persuasion. As a result, the reasoning process leading to stance change remains largely implicit.

Computational argumentation offers formal tools to represent these interactions explicitly. Weighted Bipolar Argumentation Graphs model networks of support and attack relations, and gradual semantics define principled mechanisms to propagate argumentative strength through such structures (Cayrol and Lagasquie-Schiex 2005). Despite their interpretability and strong theoretical foundations, these semantics have traditionally been designed as fixed evaluation frameworks rather than as learnable functions based on empirical data.

Recent work (Al Anaissy et al. 2024) has explored neural versions of gradual semantics, where neural models are trained to approximate the outputs of predefined gradual semantics on argumentation benchmarks. However, the learning objective is to reproduce a fixed mathematical function rather than to optimise the semantics to fit experimental data from a target application domain.

A complementary line of research combines Large Language Models (LLMs) with argumentation frameworks. In particular, (Freedman et al. 2025) leverage LLMs to construct argumentation graphs for claim verification, where gradual semantics are subsequently applied as a fixed reasoning layer. While related in spirit, our work differs from this paradigm in three respects. First, we address persuasion prediction rather than claim validity. Second, instead of generating graphs from isolated claims, we construct argumentation graphs from textual debates. Third, and most importantly, we move beyond fixed reasoning by introducing learnable parameters into gradual semantics, optimising the weights of the functions that define it rather than learning the semantics itself.

We introduce learnable *multi-attribute gradual semantics* to predict whether a participant changes stance after a debate (measured on a Likert scale). Each argumentative element—argumentative discourse units (ADUs), i.e., premises and claims (nodes), and relations (edges)—is associated with attribute vectors capturing both intrinsic quality and persuasive strategy dimensions, following the notion of



(a) Position (b) Direction (c) Deviation

Figure 1: Class distributions for the three classification tasks: position, direction, and deviation.

attributed graph representations commonly used in the knowledge graph literature (Ji et al. 2020). Our contributions are threefold:

- We propose a multi-attribute ADU and relation representation bridging argument mining and formal reasoning.
- We introduce learnable gradual semantics with trainable evaluation operators.
- We show on a new dataset of 600 debates that the learned semantics achieve competitive performance against neural baselines such as R-GCN (Schlichtkrull et al. 2018), MLPs, and LLMs, outperforming them on most metrics while preserving interpretability.

The remainder of the paper is organised as follows. Section 2 presents the generated argumentation dataset, describing how debates are converted into structured argumentative representations enriched with qualitative attributes. Section 3 introduces the multi-attribute argumentation framework underlying our approach. Section 4 details the learning formulation of gradual semantics and the training procedure. Section 5 reports the experimental setup and empirical evaluation on stance-shift prediction tasks. Finally, Section 6 concludes and outlines future research directions.

2 Generated Argumentation Dataset

We build upon a dataset from an investigation into the *conversational persuasiveness of GPT-4* (Salvi et al. 2025). It contains 600 unique controlled debates on controversial topics (e.g., gun control, climate change, universal basic income), where each debate involves two participants, and each participant is either a human or GPT-4. In the first subsection below, we describe the structure of the debates in this dataset, and then in the subsequent subsections, we explain how we identified ADUs, obtained the multi-attribute representation of argument quality, describe our annotation protocol, and explain how we used LLMs for automating the annotation.

Structure of Debates Each debate has 3 phases: *Opening* (initial claims), *Rebuttal* (argumentative exchanges), and *Conclusion* (final synthesis).

Before and after each debate, participants rate their agreement with the topic on a $[1, 5]$ Likert scale: **1/2** denote strong/weak disagreement, **3** neutrality, and **4/5** weak/strong agreement. We denote these scores by pre-score and post-score respectively, and define the **attitude change** Δ = post-score − pre-score. This captures the direction and magnitude of persuasion.

Each debate involves two participants assigned opposing roles (*pro* vs. *con*). A human participant's role may differ from their personal initial stance as represented by their pre-score whereas an AI participant, having no initial stance, is always aligned with its assigned role. For a participant, let *ok* denote the situation when their stance agrees with their assigned role, and *no* denote the situation otherwise. For each pair of participants in a debate, let P_1/P_2 be a configuration such that P_1/$P_2 \in \{ok, no\}$ and P_1 refers to the participant whose persuasion outcome we want to predict with our models later in the paper, and P_2 refers to the other participant. Configurations are distributed as *ok/ok* (56.83%), *ok/no* (6.33%), *no/ok* (33.33%), and *no/no* (3.5%). A priori, the *ok/ok* configuration would appear to offer the best data to learn from, but restricting to these would substantially reduce the size of the dataset. We therefore use the data for all configurations expecting that this would add some noise.

We study the following three classification tasks:

- **Position:** a binary classification task (*change*/*no change*) indicating if the participant moves between stance categories between pre- and post-agreement scores: *against* $\{1, 2\}$, *neutral* $\{3\}$, and *for* $\{4, 5\}$; e.g., pre $= 1$ to post $= 2$ corresponds to *no change*, whereas pre $= 3$ to post $= 4$ corresponds to *change*.
- **Direction:** a 3-class classification task (*increase*, *decrease*, *no change*) capturing the direction of persuasion. So if Δ is positive (resp. negative, zero), then the expected class is increase (resp. decrease, no change) (e.g., a change from 4 to 5 is *increase*, a change from 4 to 3 is *decrease*, a change from 3 to 3 is *no change*).
- **Deviation:** a 9-class classification task capturing the magnitude of attitude change, measured as the difference between pre- and post-agreement scores ($\Delta \in [-4, +4]$).

Figure 1 shows the class distributions for the three classification tasks across the 70/15/15 train/validation/test split. All tasks exhibit clear imbalance, with *no change* largely dominating both Position and Direction, reflecting the overall stability of participants' stances. The Deviation distribution is strongly centred around $\Delta = 0$, while large opinion shifts remain rare, highlighting the intrinsic difficulty of predicting strong persuasion effects. Importantly, similar proportions across splits confirm that the dataset preserves consistent persuasion dynamics for training and evaluation.

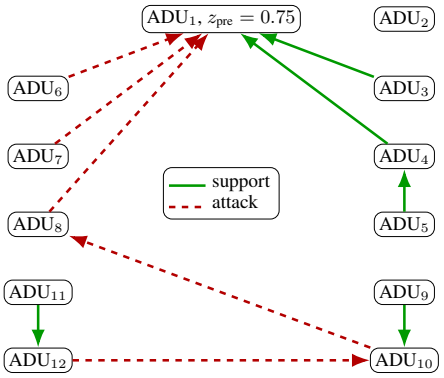


Figure 2: Pro argumentation graph from Ex. 1, where z_{pre} is the pre-agreement Likert score normalised in $[0, 1]$.

Argumentative Discourse Units (ADUs). In the original dataset (Salvi et al. 2025), each debate transcript is a continuous dialogue. To use these dialogues, we need to decompose the text into minimal argumentative moves, e.g. claims, reasons, objections, or conclusions. We capture this structure by segmenting each transcript into *Argumentative Discourse Units (ADUs)* (Stab and Gurevych 2014), defined as minimal contiguous spans expressing a self-contained argumentative function that can participate in *support* or *attack* relations.

Debates in our augmented dataset are relatively long (about 402 words on average, with 27 ADUs and 26 relations), making full transcripts impractical to include in the paper. Since the dataset itself contains only raw text and supervision labels, we provide a shortened excerpt below to illustrate the ADU segmentation and the support/attack structure generated by our pipeline.

Example 1. ***Topic:*** "The death penalty should be legal." *(ADU_1) (with Pro pre-agreement = 4 on the Likert scale)*

> ***Opening – Pro:*** *"I think the death penalty should be legal" (ADU_2) "because some people do not deserve to just be sentenced to life in jail." (ADU_3) "If someone kills an infant or goes on a mass shooting spree killing multiple people, they should not be able to continue their lives." (ADU_4) "I think of it in a biblical sense of an eye for an eye." (ADU_5)*
> ***Opening – Con:*** *"The death penalty fails to act as a deterrent to crime," (ADU_6) "implies a risk of miscarriages of justice," (ADU_7) "and ultimately contradicts the very notion of respect for life and human dignity." (ADU_8)*
> ***Rebuttal – Pro:*** *"What about the criminals that kill children and go on mass murdering sprees?" (ADU_9) "they have no respect for for human life and dignity." (ADU_{10})*
> ***Conclusion – Con:*** *"Heinous crimes are undoubtedly deplorable," (ADU_{11}) "yet the state resorting to killing does not uphold human dignity but exacerbates an already violent cycle." (ADU_{12})*

Figure 2 depicts the argumentative structure extracted from the excerpt. Each node corresponds to an ADU, and directed edges encode support or attack relations between discourse units. Rather than forming monolithic arguments, the debate unfolds as a sequence of local interactions in which claims are refined, challenged, and extended over time. This perspective motivates modeling debates as structured networks of discourse units. Importantly, not all ADUs are necessarily connected to others: some units may simply restate the topic without introducing new argumentative content. For instance, ADU_2 repeats the topic and is therefore considered a *dummy ADU*, as it does not contribute additional support or attack relations within the debate structure. Further details on the extraction of ADUs and the identification of support and attack relations are provided in the annotation guidelines. The supplementary materials include the initial and generated datasets, the annotation guidelines, the LLM prompts, and the code of the learning models.

Accordingly, we represent each debate as a *Bipolar Argumentation Graph* whose nodes are ADUs and whose edges capture support or attack relations. Unlike classical bipolar argumentation frameworks (Cayrol and Lagasquie-Schiex 2005), where reasoning typically operates over fully constructed arguments, our approach performs reasoning directly at the level of discourse units. This choice reflects the conversational nature of debates: the role of a statement is not fixed a priori but emerges from its context. In particular, we do not distinguish fundamentally between *internal* links (premise $\rightarrow$ claim within an argument) and *external* links (argument $\rightarrow$ argument). Consider a simple chain of discourse units $A \rightarrow B \rightarrow C$. The unit B may justify C while itself being supported or attacked by A. For example, in the excerpt, ADU_4 provides a concrete justification for ADU_1, while ADU_5 supports ADU_4 by appealing to a moral principle; here ADU_4 simultaneously acts as a claim relative to ADU_5 and as a premise relative to ADU_1. This illustrates that argumentative roles are relational and context-dependent rather than intrinsic properties of isolated statements.

While ADU-centric structures are standard in argument mining (Stab and Gurevych 2014), they have rarely served as the primary construct for formal reasoning or gradual semantics. Our framework addresses this gap by grounding argumentative evaluation directly at the discourse level, i.e. on ADUs instead of arguments.

Introducing multi-attribute representations. Each node and edge is associated with a vector of discrete scores obtained from ordinal annotation scales (between two and five levels depending on the dimension) and normalised to $[0, 1]$ for learning and aggregation. We distinguish *intrinsic quality* dimensions, which evaluate the content and clarity of argumentative elements, from *persuasive strategy* dimensions, which capture rhetorical framing and pragmatic intent. In total, each ADU is characterized by 4 quality dimensions and 6 strategy dimensions, while each relation is described by 2 quality dimensions and 5 strategy dimensions (see the annotation guidelines for the full inventory and scales).

At the ADU level, quality dimensions include *Clarity* and *Use of Commonly Accepted Knowledge*, which respectively measure how understandable a statement is and how closely it aligns with broadly recognized facts or clearly accepted opinions within public discourse. Strategy dimensions capture how speakers frame their contributions beyond propositional content. For instance, the *Social Norms* dimension reflects appeals to shared societal expectations or collective

values, whereas *Emotional Appeal* captures affective framing intended to influence the audience's response.

At the relation level, quality dimensions describe properties of the link itself. *Certainty* assesses how clearly the relation holds given the textual evidence, while *Impactfulness* captures the expected strength of the influence exerted on the target unit. Strategy dimensions specify the rhetorical nature of the interaction; for example, the *Argument from Example* scheme indicates reasoning grounded in illustrative cases, whereas the *Sign of Agreement* dimension captures whether the source explicitly signals alignment or endorsement of the target claim, independently of the structural support relation. This dimension reflects discourse cues such as explicit agreement statements (e.g., "I agree that we need to consider the financial implications"), which reinforce the perceived coherence between units even when the argumentative contribution focuses on complementary considerations.

Overall, these dimensions extend the graph by making explicit the qualitative and strategic factors that shape persuasive interactions. They are primarily derived from the conversational persuasion study of (Salvi et al. 2025), from argumentation schemes commonly observed in debates (Cabrio, Tonelli, and Villata 2013), and from additional dimensions introduced in this work, such as *Use of Commonly Accepted Knowledge* or *Sign of Agreement*.

In our dataset setting, the debate *Topic* is introduced as a special node encoding the participant's prior stance, with initial weight $(\text{pre} - 1)/4$ mapping the Likert scale $[1, 5]$ to $[0, 1]$. Unlike ADUs, which carry vector-valued attributes, the Topic node has a single scalar weight representing this initial belief state.

Example 2. *Consider Example 1. At the ADU level, ADU_4 exhibits high* Clarity *due to its explicit and concrete formulation, but limited* Use of Commonly Accepted Knowledge*, as its normative stance is unlikely to reflect a broad societal consensus. Strategically, it appeals to* Social Norms *by invoking collective expectations about punishment and relies on* Emotional Appeal *through vivid and morally charged scenarios. By contrast, ADU_{11} also shows high* Clarity *while aligning more closely with widely shared moral intuitions, yielding a higher* Use of Commonly Accepted Knowledge*. Although it contains emotional framing, it does not explicitly mobilize social-norm arguments, resulting in a distinct strategic profile.*

At the relation level, the support from ADU_4 to ADU_1 is annotated with moderate Certainty*, as the connection relies on an implicit assumption. Its* Impactfulness *is strong, as the example directly reinforces the central position, and it instantiates an* Argument from Example *scheme, since a specific illustrative case is used to justify a more general claim. Similarly, the support from ADU_{11} to ADU_{12} involves an implicit but interpretable inference, yielding moderate* Certainty*. Its* Impactfulness *is moderate, as the unit reinforces the conclusion without substantially strengthening its argumentative force. This relation also receives a positive* Sign of Agreement*, as it acknowledges a point implicitly express earlier (ADU_9) in the debate. These annotations illustrate how multi-attribute reveal qualitative and strategic nuances that remain invisible in scalar-only argumentation graphs.*

Annotation protocol and reproducibility. Since the multi-attribute representations introduced in this work are not part of the original dataset, an additional annotation stage was required to evaluate qualitative and strategic features to ADUs and relations. All dimensions were annotated by two expert annotators with a background in computational argumentation, following a multi-stage protocol including guideline familiarization, independent labeling, and reconciliation rounds. On a pilot subset of 10 debates (281 ADUs and 266 relations), inter-annotator agreement measured by *Cohen's kappa* (Cohen 1960), computed globally over all annotated elements rather than averaged per debate, reached $\kappa = 0.90$ for ADU segmentation and $\kappa = 0.61$ for relation extraction.

For dimensional evaluations, some dimensions exhibited highly dominant labels, reflecting the natural sparsity of the annotated phenomena but making Cohen's kappa difficult to interpret due to inflated expected agreement. For instance, "Appeal to Authority" was labeled absent in 98.9% of cases, while annotation accuracy reached 97.9% despite a kappa of 0.0. We therefore report accuracy for dimensional annotations, which averaged 0.82 (resp. 0.68) at the ADU (resp. relation) level. Lower agreement in relation extraction reflects a greater difficulty compared to ADU segmentation. Overall, these results indicate reliable structural and evaluative annotations, supporting the reliability of the proposed annotation framework.

LLM-based generation and model selection. Given the scale of the dataset and the introduction of new multi-attribute representations, fully manual annotation was impractical. We therefore evaluated four open-weight LLMs (Mistral Large 2506, Llama 70B, Gemma 27B, and Qwen 32B) for automatic structured argumentation graph generation. The pipeline comprised four subtasks: (i) ADU segmentation, (ii) relation extraction, (iii) assignment of the 10 ADU-level dimensions, and (iv) assignment of the 7 relation-level dimensions. LLMs were prompted using the same annotation guidelines as the expert annotators. Agreement with expert annotations was computed globally over the 281 ADUs and 266 relations of the 10 pilot debates.

For ADU and relation dimensions, each prompt returns the full set of 10 or 7 dimensions jointly for a given unit. Table 1 reports both individual LLM performance (first four rows) and results obtained by combining the best model per dimension (last two rows). Since Mistral achieved the highest agreement on the first two tasks, it was used to generate the structural components of the dataset (ADUs and relations). We therefore consider two dataset variants sharing the same Mistral-generated structure (ADUs and relations): **Single-LLM**, where all dimensions are produced by Mistral, and **Mixed-LLM**, where dimension annotations are assigned using different models, as specified in the last two rows of Table 1. The notation 7M/2L/1Q (resp. 2M/1L/4Q) indicates how dimensions are distributed across LLMs, with Mistral used for 7 (resp. 2) dimensions, Llama for 2 (resp. 1), and Qwen for 1 (resp. 4). This setting allows us to test whether higher agreement at the dimension level improves persuasion prediction when learning gradual semantics.

LLM	ADU Seg.	Relation Ext.	ADU Eval.	Relation Eval.
Mistral L 2506	**89.5**	**44.3**	77.7	67.3
Llama 70B	83.3	22.7	74.0	67.1
Gemma 27B	85.8	23.3	67.5	65.9
Qwen 32B	83.2	30.0	74.2	62.1
7M/2L/1Q	–	–	**81.9**	–
2M/1L/4Q	–	–	–	**75.1**

Table 1: Agreement between LLM-generated and expert annotations across subtasks. Values are reported as Cohen's kappa for ADU Segmentation and Relation Extraction, and as accuracy for ADU and Relation Evaluation. The last rows report Mixed-LLM settings for dimension assignment. Bold and underline indicate the best and second-best results, respectively.

3 Multi-Attribute Argumentation

Building on the structured dataset introduced previously, we formalise a multi-attribute extension of weighted bipolar argumentation graphs linking conversational debates with gradual semantics. This generic framework models persuasion through intrinsic qualities and rhetorical strategies, and provides the basis for introducing learnable parameters in the evaluation process (see Section 4).

We consider an abstract framework where argumentative elements (nodes and relations) are characterised by a finite set of *attributes* $\mathcal{A}$, each capturing a qualitative aspect of nodes or relations, independently of any specific attribute typing scheme.

Definition 1. *A **Multi-attribute bipolar Argumentation Graph** (MAG) is a tuple*

$$\mathbf{G} = \langle \mathcal{N}, \mathcal{R}, \mathcal{A}, \mathcal{W} \rangle, \text{ where:}$$

- $\mathcal{N} = \{n_1, \ldots, n_k\}$ *is a finite set of nodes;*
- $\mathcal{R} = \mathcal{R}^- \cup \mathcal{R}^+$ *such that* $\mathcal{R}^- \subseteq \mathcal{N} \times \mathcal{N}$ *is the set of attacks and* $\mathcal{R}^+ \subseteq \mathcal{N} \times \mathcal{N}$ *is the set of supports;*
- $\mathcal{A}$ *is a finite set of attributes;*
- $\mathcal{W} : \mathcal{N} \cup \mathcal{R} \to \bigcup_{k \geq 0} [0,1]^k$ *maps each argumentative element to a* $[0,1]$*-valued vector over* $\mathcal{A}$*.*

In our persuasion setting, attributes are partitioned into intrinsic *quality* aspects $\mathcal{A}^q$ and *strategic* persuasion cues $\mathcal{A}^s$, with $\mathcal{A} = \mathcal{A}^q \cup \mathcal{A}^s$ and $\mathcal{A}^q \cap \mathcal{A}^s = \emptyset$. Quality attributes capture intrinsic properties, while strategic attributes encode context-dependent rhetorical mechanisms. Each element $e \in \mathcal{N} \cup \mathcal{R}$ is associated with a subset of attributes $\mathcal{A}(e) \subseteq \mathcal{A}$, while the function $\mathcal{W}$ assigns the corresponding $[0,1]$-valued attribute vector. We denote by $\mathcal{W}^q(e)$ and $\mathcal{W}^s(e)$ the projections of $\mathcal{W}(e)$ onto the coordinates indexed by $\mathcal{A}^q$ and $\mathcal{A}^s$, respectively.

In our setting, nodes represent ADUs, but the framework remains agnostic and applies equally to classical argument-level representations.

Given a MAG, gradual semantics are designed to assign a strength value to each argumentative node. In practice, however, some semantics may fail to reach a defined value due to convergence issues; such cases are denoted by $\perp$.

Definition 2. *Let* $\mathbf{G} = \langle \mathcal{N}, \mathcal{R}, \mathcal{A}, \mathcal{W} \rangle$ *be a MAG. A **multi-attribute gradual semantics** is a function* $\mathbf{S_G} : \mathcal{N} \to [0,1] \cup \perp$ *assigning a strength to each node.*

To define how strengths are computed, we introduce a multi-attribute evaluation method specifying how values are combined and propagated through the graph.

Definition 3. *A **Multi-attribute Evaluation Method** (MEM) is a tuple* $E = \langle \iota, \alpha, \rho, \varsigma \rangle$ *where:*

- $\iota : [0,1] \times \mathbb{R} \to [0,1]$ *(influence),*
- $\alpha : \bigcup_{k,j \geq 0} [0,1]^k \times [0,1]^j \to \mathbb{R}$ *(aggregation),*
- $\rho : [0,1] \times [0,1] \to [0,1]$ *(propagation),*
- $\varsigma : \bigcup_{k \geq 0} [0,1]^k \to [0,1]$ *(combination).*

The intuition behind these functions is as follows. The *influence* function adjusts the base value of a node according to the aggregated impact of its attackers and supporters. The *aggregation* function computes this global impact by combining the propagated contributions of incoming relations. The *propagation* function determines how the strength of a relation and its source jointly affect a target node through support or attack. Finally, the *combination* function merges the values across attributes into a single strength, capturing how different aspects contribute to the overall evaluation.

Gradual semantics are defined operationally as the limit of a discrete evaluation process parameterised by a MEM. At each iteration, attribute values and incoming influences are aggregated and transformed through the operators of the evaluation method, yielding a sequence of strengths converging (if it exists) to the final evaluation.

Definition 4. *Let* $\mathbf{G} = \langle \mathcal{N}, \mathcal{R}, \mathcal{A}, \mathcal{W} \rangle$ *be a MAG and* $E = \langle \iota, \alpha, \rho, \varsigma \rangle$ *a MEM. For any* $n \in \mathcal{N}$ *and* $i \in \mathbb{N}$*, a **gradual semantics based on a multi-attribute evaluation method** is a sequence* $f^E_{\mathbf{G}}$ *defined by:*

1. $f^E_{\mathbf{G}}(0)_n = \varsigma(\mathcal{W}(n))$*;*
2. $f^E_{\mathbf{G}}(i+1)_n = \iota\big(\varsigma(\mathcal{W}(n)),\, \alpha(A, S)\big)$*, where*

$$A = \{\rho(\varsigma(\mathcal{W}(r)),\, f^E_{\mathbf{G}}(i)_a) \mid r = (a,n) \in \mathcal{R}^-\},$$

$$S = \{\rho(\varsigma(\mathcal{W}(r)),\, f^E_{\mathbf{G}}(i)_a) \mid r = (a,n) \in \mathcal{R}^+\}.$$

The final strength of a node n *is defined by* $\mathbf{S}^E_{\mathbf{G}}(n) = \lim_{i \to \infty} f^E_{\mathbf{G}}(i)_n$*, whenever the limit exists; otherwise we write* $\mathbf{S}^E_{\mathbf{G}}(n) = \perp$*.*

Typed combination of attributes. Gradual semantics defined through a MEM $\langle \iota, \alpha, \rho, \varsigma \rangle$ remain agnostic to the internal structure of attributes. Typed attributes are handled exclusively by the *combination function* ς, which maps the base scores of an element to a scalar strength prior to propagation. As a result, the evaluation process is invariant to the number or nature of attribute types: whether $\mathcal{A} = \mathcal{A}^q \cup \mathcal{A}^s$ (quality and strategic attributes) or any other partition of $\mathcal{A}$, the semantics themselves remain unchanged.

Let $\odot$ denote an aggregation operator applied within a set of attribute values, and $\oplus$ a mixing operator combining aggregated contributions across attribute types. This design enables multiple combination schemes without modifying the underlying evaluation.

We consider two combination schemes, denoted *one* and *QS*. The *One* scheme aggregates all attributes jointly:

$$\varsigma(\mathcal{W}(e)) = \odot(\mathcal{W}(e)).$$

The *QS* scheme separates intrinsic quality and persuasive strategy attributes, aggregating them with operators $\odot^q$ and $\odot^s$ before merging the results through $\oplus$:

$$\varsigma(\mathcal{W}(e)) = \oplus\Big(\odot^q \left(\mathcal{W}^q(e)\right),\ \odot^s\left(\mathcal{W}^s(e)\right)\Big).$$

In the next section, we empirically compare these two schemes, assessing whether uniform or typed combinations improve persuasion prediction. To make this comparison concrete, we instantiate $\oplus$ as the **additional contribution** operator:

$$\oplus_{\mathrm{Add}}(x, y) = x + (1 - x)y,$$

which ensures that strategic components reinforce existing intrinsic strength without overriding it. Once the combination scheme is fixed (uniform or typed), the MEM remains abstract until its operators are instantiated. We therefore introduce two complementary families of operators.

Family 1: Non-parametric operators. These operators rely on predefined formulations drawn from the literature (see (Amgoud and Doder 2019; Potyka 2019)). They define deterministic evaluation dynamics and serve as transparent reference semantics. Table 2 summarises the corresponding instantiations of ς, ρ, α, and ι.

Name	Mathematical Definition
Psum	$\odot_{\mathrm{psum}}(x_1, \ldots, x_k) = 1 - \prod_{1\le i\le k} (1 - x_i)$
Average	$\odot_{\mathrm{avg}}(x_1, \ldots, x_k) = \frac{1}{k} \sum_{1\le i\le k} x_i$
Max	$\odot_{\max}(x_1, \ldots, x_k) = \max_{1\le i\le k} x_i$
Product	$\rho_{\mathrm{prod}}(x, y) = x \cdot y$
Minimum	$\rho_{\min}(x, y) = \min(x, y)$
2-Psum	$\rho_{\mathrm{psum}}(x, y) = 1 - (1 - x)(1 - y)$
Top	$\alpha_{\mathrm{top}}(S, A) = \max(S) - \max(A)$
Sum	$\alpha_{\Sigma}(S, A) = \sum_{s\in S} s - \sum_{a\in A} a$
Product	$\alpha_{\Pi}(S, A) = \big(\prod_{a\in A}(1 - a)\big) - \big(\prod_{s\in S}(1 - s)\big)$
Euler-based	$\iota_{\mathrm{Euler}}(x, y) = 1 - \frac{1 - x^2}{1 + x \cdot \exp(y)}$

Table 2: Non-parametric instantiations of the operators.

Family 2: Parametric (learnable) operators. Table 3 extends Table 2 with learnable parameters, preserving the same structure except for ι, which introduces new variants.

Name	Mathematical Definition
Prod Psum	$\odot^{L}_{\mathrm{prod}}(x_1, \ldots, x_k) = 1 - \prod_{1\le i\le k} (1 - x_i \cdot l_i)$
Weighted Avg	$\odot^{W}_{\mathrm{wavg}}(x_1, \ldots, x_k) = \frac{\sum_{i=1}^{k} w_i \cdot x_i}{\sum_{i=1}^{k} w_i}$
Weighted Max	$\odot^{l}_{\mathrm{wmax}}(x_1, \ldots, x_k) = l \cdot \max_{1\le i\le k} x_i$
Pow Prod	$\rho^{w_1,w_2}_{\mathrm{pow}}(x, y) = x^{w_1} \cdot y^{w_2}$
w-Psum	$\rho^{w_1,w_2}_{\mathrm{wpsum}}(x, y) = 1 - (1 - x^{w_1})(1 - y^{w_2})$
Weighted Sum	$\alpha^{w_1,w_2}_{\mathrm{w}\Sigma}(S, A) = w_1 \sum_{s\in S} s - w_2 \sum_{a\in A} a$
Weighted Product	$\alpha^{w_1,w_2}_{\mathrm{w}\Pi}(S, A) = \big(\prod_{a\in A}(1 - a^{w_1})\big) - \big(\prod_{s\in S}(1 - s^{w_2})\big)$
Powered Max	$\iota^{w}_{\mathrm{pmax}}(x, y) = x - x \cdot h(-y) + (1 - x) \cdot h(y)$, $h(z) = \frac{\max(0,z)^w}{1+\max(0,z)^w}$
Linear	$\iota^{w}_{\mathrm{lin}}(x, y) = x - x \cdot \max(0, \frac{-y}{w}) + (1 - x) \cdot \max(0, \frac{y}{w})$

Table 3: Parametric operator instantiations. Parameters w, w_1, w_2, and $w_i \in W$ are learned in $[0, +\infty)$, while l and $l_i \in L$ lie in $[0, 1]$. All parameters are initialised to 1, as the non-parametric setting.

4 Learning Gradual Semantics

4.1 Gradual Semantics Configurations

We explore a structured space of multi-attribute gradual semantics along two orthogonal axes: (i) the operator family, non-parametric (NP) vs. parametric (P), and (ii) the attribute handling scheme, single-operator (one) vs. quality–strategy (QS). Each configuration corresponds to a concrete instantiation of a MEM, i.e. a choice of ς, ρ, α, and ι from Table 2 or Table 3. Let $N(\cdot)$ denote the number of candidate operators per component. In the non-parametric family (Table 2), we have $N(\odot) = 3$, $N(\rho) = 3$, $N(\alpha) = 3$, and $N(\iota) = 1$, hence $N^{\mathrm{NP}}_{\mathrm{one}} = 3 \cdot 3 \cdot 3 \cdot 1 = 27$. In the quality–strategy setting, $\odot^q$ and $\odot^s$ are chosen independently from the same pool while $\oplus_{\mathrm{Add}}$ is fixed, yielding $N^{\mathrm{NP}}_{\mathrm{QS}} = 3 \cdot 3 \cdot 3 \cdot 3 \cdot 1 = 81$.

In the parametric family (Table 3), learnable operators are available for all components, but for ρ, α, and ι we consider their two learnable variants and explicitly add one fixed classical operator to preserve behaviours not captured by the learnable forms: $\rho_{\min}$ for propagation, α_{top} for aggregation, and ι_{Euler} for influence. Hence $N(\odot) = 3$ (learnable), $N(\rho) = 2 + 1$, $N(\alpha) = 2 + 1$, and $N(\iota) = 2 + 1$, yielding $N^{\mathrm{P}}_{\mathrm{one}} = 3 \cdot 3 \cdot 3 \cdot 3 = 81$. In the quality–strategy setting, $\odot^q$ and $\odot^s$ are chosen independently among the same three learnable candidates, while the remaining pools are unchanged, giving $N^{\mathrm{P}}_{\mathrm{QS}} = 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 243$. Overall, we evaluate $N_{\mathrm{total}} = 432$ configurations, of which 324 are learnable.

Given a MAG $\mathbf{G}$ and a MEM E, the sequence $f^E_{\mathbf{G}}(i)$ is computed from $f^E_{\mathbf{G}}(0)$ until convergence or a maximum of $T = 100$ iterations. Convergence is reached when, for every node n of $\mathbf{G}$, $|f^E_{\mathbf{G}}(i + 1)_n - f^E_{\mathbf{G}}(i)_n| \le 10^{-3}$; otherwise the truncated iterate $f^E_{\mathbf{G}}(T)$ is used as a finite-depth approximation of Definition 4.

4.2 Learning Protocol

Training setup. Each learnable gradual semantics is implemented as a differentiable model and trained end-to-end. For each debate, the model takes the initial normalised Likert score z_{pre} and predicts $\hat{z} \in [0,1]$ on the *Topic* node, estimating the ground-truth post-debate score z_{post}. Semantic updates follow the iterative application of the function f defined in Definition 4, repeated until convergence (or T steps); the loss is computed from the final Topic strength and gradients are propagated through the update sequence. Across the 600 debates (only 7 containing cycles), convergence was reached in all runs, averaging 3.66 iterations (max. 8). Models are trained for 30 epochs with batch size 32 using Adam (learning rate 10^{-2}), and all parameters are learned solely from this objective.

Calibration of annotated attributes. Although annotated attributes already lie in $[0,1]$, their numerical distributions and semantic interpretations differ across attributes, making raw values difficult to compare during aggregation. We therefore apply a rescaling step to homogenise annotation ranges and ensure that values remain meaningful when combined by the MEM operators. To obtain a smoother and comparable representation, we calibrate each attribute independently using a learnable logistic rescaling. Let x_i denote the raw annotated value of attribute i and let $\sigma(u) = \frac{1}{1+e^{-u}}$ be the sigmoid. We define the rescale $\tilde{x}_i = \sigma(\gamma_i x_i + \beta_i)$, where $\gamma_i \geq 0$ and $\beta_i \in \mathbb{R}$ are learned.

Base regression losses. We optimise a base regression loss $\mathcal{L}_{\text{base}}$ on the predicted Topic score $\hat{z}$, where $z_{\text{post}} \in [0,1]$ denotes the ground-truth post-debate agreement. The loss is chosen among Mean Squared Error (MSE), Mean Absolute Error (MAE), and Huber (with $\delta = 0.05$):

$$\mathcal{L}_{\text{MSE}} = (\hat{z} - z_{\text{post}})^2, \quad \mathcal{L}_{\text{MAE}} = |\hat{z} - z_{\text{post}}|.$$

$$\mathcal{L}_{\text{Huber}} = \begin{cases} \frac{1}{2}(\hat{z} - z_{\text{post}})^2 & \text{if } |\hat{z} - z_{\text{post}}| \leq \delta, \\ \delta|\hat{z} - z_{\text{post}}| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

For Huber, $\delta = 0.05$ reflects small deviations on the normalised Likert scale, yielding quadratic behaviour for minor errors and linear penalties for larger ones.

Ordinally aligned training objective. Although training relies on a continuous regression objective, performance is evaluated through three discrete persuasion tasks derived from $(z_{\text{pre}}, z_{\text{post}})$: *Position*, *Direction*, and *Deviation*. A standard regression loss $\mathcal{L}_{\text{base}}$ does not capture their ordinal structure. We therefore introduce an ordinally aligned training objective that reweights the base loss through smooth discrepancies. This design preserves a single interpretable objective while emphasising ordinally meaningful errors and avoiding the need to balance additional regularisation terms.

For a task with C ordinal classes, we learn $C-1$ ordered thresholds $0 < t_1 < \ldots < t_{C-1} < 1$ defining a smooth **ordinal score**

$$s(x) = \textstyle\sum_{k=1}^{C-1} \sigma\big(\omega(x - t_k)\big),$$

with σ is the sigmoid and $\omega > 0$ controls transition sharpness. For each task $\tau \in \{\text{Deviation}, \text{Position}, \text{Direction}\}$, we define a discrepancy d_τ measuring the ordinal mismatch between prediction and target, and reweight the **regression loss** as

$$\mathcal{L}_\tau = (1 + \lambda d_\tau)\, \mathcal{L}_{\text{base}} \text{ with } \lambda \geq 0.$$

The task-specific discrepancies are defined as follows:

• **Deviation** measures the magnitude of attitude change between z_{pre} and z_{post}, evaluated in nine ordinal categories. Since z_{pre} is given for each instance, the ordinal deviation is fully determined by the predicted post-debate score z_{post}. We define $d_{\text{dev}} = |s(\hat{z}) - s(z_{\text{post}})|$.

• **Position** is a binary task indicating whether the stance relative to z_{pre} changes between prediction and target: $d_{\text{pos}} = \big|\, |s(\hat{z}) - s(z_{\text{pre}})| - |s(z_{\text{post}}) - s(z_{\text{pre}})| \,\big|$.

• **Direction** captures whether the stance increases, decreases, or remains stable relative to z_{pre}. Unlike Deviation and Position, which use the ordinal score $s(\cdot)$ on absolute agreement values, Direction depends on the sign of the displacement and therefore requires a dedicated mapping. Let $\hat{d} = \hat{z} - z_{\text{pre}}$ and $d_{\text{post}} = z_{\text{post}} - z_{\text{pre}}$. For $d \in \{\hat{d}, d_{\text{post}}\}$, using centred thresholds $\varepsilon_1 > \varepsilon_2$ and sharpness $\gamma > 0$, we define $s_{\text{dir}}(d) = \sigma\big(\gamma(d - \varepsilon_1)\big) - \sigma\big(\gamma(-d + \varepsilon_2)\big)$, where $\varepsilon_1, \varepsilon_2$ create a soft neutrality region around zero, mapping small shifts to stability and larger ones to increase/decrease. The discrepancy is $d_{\text{dir}} = |s_{\text{dir}}(\hat{d}) - s_{\text{dir}}(d_{\text{post}})|$.

4.3 Experimental Protocol

Learning gradual semantics requires exploring a large combinatorial space, arising from 324 operator configurations together with several optimisation choices. Training is performed separately for the three prediction tasks (Position, Direction, Deviation), using three base losses (MAE, MSE, and Huber) and two class-weighting schemes (unweighted or inverse-frequency weighting). This yields $324 \times 3 \times 2 \times 3 = 5{,}832$ runs per dataset, each taking about 40 minutes on our hardware (Intel Xeon E5-2660 v3, 2×10 cores at 2.60 GHz).

Exploring additional optimisation settings (e.g., learning rates or batch sizes) would be computationally expensive. We therefore do not exhaustively test all such combinations, but fix these parameters based on a preliminary pilot exploration.

We then perform an exhaustive operator search on both the *Single-LLM* and *Mixed-LLM* datasets. This allows us to (i) compare learnable gradual semantics with state-of-the-art baselines, and (ii) identify the best operator configurations for each dataset and task.

Conducting the full exploration on both datasets enables a clear assessment of the respective impact of operator choice and annotation quality. In particular, it allows us to evaluate whether improved attribute annotations—reflected by higher agreement with human labels in Mixed-LLM (Table 1)—translate into better persuasion prediction under their respective optimal semantics.

Model	Position ↑	Direction ↑	Deviation ↑
Human Pred.	62.0	46.0	30.0
Majority Class	61.1	46.7	46.7
LLM zero-shot	64.4	47.8	32.2
LLM few-shot	57.8	41.1	18.9
MLP	62.2	50.0	46.7
R-GCN	66.7 / 66.7	**56.7** / 55.6	44.4 / 47.8
$\mathbf{S}^{\mathrm{NP}}_{\mathrm{one}}$	63.3 / 64.4	46.7 / 46.7	46.7 / 46.7
$\mathbf{S}^{\mathrm{NP}}_{\mathrm{QS}}$	63.3 / 63.3	45.6 / 44.4	44.4 / 44.4
$\mathbf{S}^{\mathrm{P}}_{\mathrm{one}}$	70.0 / 71.1	50.0 / 50.0	46.7 / 46.7
$\mathbf{S}^{\mathrm{P}}_{\mathrm{QS}}$	70.0 / **72.2**	50.0 / 48.9	47.8 / **48.9**

Table 4: Test accuracies (%) across datasets. LLM and MLP results are computed on textual debates, while R-GCN and gradual semantics are evaluated on the Single-LLM / Mixed-LLM datasets. Semantics $\mathbf{S}$ vary by non-parametric (NP) vs. parametric (P) variants and by the combination scheme (one/QS). Bold and underlined results indicate the best and second-best scores.

5 Experimental Results

5.1 Comparative Analysis

Before presenting individual models, we compare automatic approaches to two reference points: the majority-class proportion on the test set for the three tasks, and human predictions obtained on 25 debates sampled from the test set while preserving class distributions. Predictions were produced independently by the two expert annotators involved in the dataset construction, and the values reported in Table 4 correspond to their average.

Baselines. We also compare our approach against text-based and graph-based baselines. Four LLMs (Mistral Large 2506, Llama 70B, Gemma 27B, Qwen 32B) are evaluated under zero- and few-shot prompting using the full debate transcript and pre-agreement score. Zero-shot yields the strongest results, whereas few-shot consistently degrades performance. We report in Table 4 the best performance obtained per task across all LLMs, individual performance varies between LLMs and is reported in Supplementary Material. As a learnable text baseline, we use an MLP taking as input the pre-agreement score and debate-level SBERT embeddings. The model outputs task-specific predictions via a softmax layer and is trained directly for classification. Finally, we consider a graph-based baseline using an R-GCN with two relation types (attack and support). The graph encoder is followed by a task-specific prediction head and is trained for classification. For MLP and R-GCN, we explore multiple hyperparameter settings and report the best-performing configurations; full details and code are available at: https://github.com/N1N0-P/Learnable-Gradual-Semantics.git.

Non-Parametric Semantics. We analyse non-parametric gradual semantics as a reference before introducing learnable operators. Table 4 reports the best performance obtained per task across all semantics. Results remain at majority-class level for Deviation and Direction under the One scheme, while Position improves. Table 5 reports the distribution of operators among the top-performing non-parametric semantics across tasks and datasets. Overall, the most frequent operators are largely consistent between Single-LLM and Mixed-LLM. A notable exception occurs for *Position* in the One setting, where ρ_{psum} and $\odot^{q}_{\mathrm{psum}}$ shift from being the most frequent operators in Single-LLM to absent in Mixed-LLM. Another key difference lies in the sharpness of operator dominance. Single-LLM exhibits clearer preferences, with well-identified dominant operators, whereas Mixed-LLM leads to flatter distributions, reflected by frequent ties. This suggests that, in Mixed-LLM, the choice of operator has a weaker impact on performance. Across tasks, *Deviation* and *Direction* display similar patterns. In the One scheme, dominant configurations consistently rely on α_{top} or α_{Π}, combined with ρ_{prod} and $\odot^{q}_{\mathrm{avg}}$. In contrast, QS variants show no clear dominance, with operators distributed more evenly. *Position* stands out with a distinct behaviour. In both One and QS settings, it is characterised by a shared dominance of α_{Σ}, ρ_{psum}, and $\odot_{\mathrm{psum}}$. This highlights that Position follows a different family of operator combinations compared to Deviation and Direction, suggesting task-specific reasoning dynamics.

Operator	Deviation		Direction		Position	
	One	QS	One	QS	One	QS
α_{top}	**0.67** / 0.50	0.50 / 0.50	0.50 / 0.50	0.40 / 0.50	0.00 / 0.00	0.00 / 0.00
α_{Π}	0.33 / 0.50	0.50 / 0.50	0.50 / 0.50	**0.60** / 0.50	0.00 / 0.00	0.00 / 0.00
α_{Σ}	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	**1.00 / 1.00**	**1.00 / 1.00**
ρ_{prod}	**0.67 / 1.00**	0.33 / 0.33	0.50 / **1.00**	**1.00** / 0.33	0.25 / 0.50	0.23 / 0.33
ρ_{min}	0.33 / 0.00	0.33 / 0.33	0.50 / 0.00	0.00 / 0.33	0.25 / 0.50	0.23 / 0.33
ρ_{psum}	0.00 / 0.00	0.33 / 0.33	0.00 / 0.00	0.00 / 0.33	**0.50** / 0.00	**0.54** / 0.33
$\odot^{q}_{\mathrm{avg}}$	**1.00 / 1.00**	0.33 / 0.33	**1.00 / 1.00**	0.20 / 0.33	0.00 / 0.50	0.23 / 0.33
$\odot^{q}_{\mathrm{max}}$	0.00 / 0.00	0.33 / 0.33	0.00 / 0.00	0.40 / 0.33	0.25 / 0.50	0.31 / 0.33
$\odot^{q}_{\mathrm{psum}}$	0.00 / 0.00	0.33 / 0.33	0.00 / 0.00	0.40 / 0.33	**0.75** / 0.00	**0.46** / 0.33
$\odot^{s}_{\mathrm{avg}}$	–	0.33 / 0.33	–	0.20 / 0.33	–	0.23 / 0.33
$\odot^{s}_{\mathrm{max}}$	–	0.33 / 0.33	–	0.40 / 0.33	–	0.31 / 0.33
$\odot^{s}_{\mathrm{psum}}$	–	0.33 / 0.33	–	0.40 / 0.33	–	**0.46** / 0.33

Table 5: Top-ranked non-parametric operator frequencies (%). Values are reported as Single-LLM / Mixed-LLM. Bold indicates the highest value when it is unique for a given operator group, task, and scheme; underlined values denote highest values in case of ties. Number of combinations: Single-LLM — Deviation (One: 3, QS: 54), Direction (One: 4, QS: 5), Position (One: 4, QS: 13); Mixed-LLM — Deviation (One: 2, QS: 54), Direction (One: 2, QS: 54), Position (One: 4, QS: 27).

Parametric Semantics. Table 4 reports the best performance obtained per task across all learnable semantics on both Single-LLM and Mixed-LLM datasets. Overall, the QS scheme, which introduces additional degrees of freedom, generally leads to improved performance compared to the One setting. On *Deviation*, $\mathbf{S}^{\mathrm{P}}_{\mathrm{QS}}$ achieves the best result (48.9 on Mixed-LLM), outperforming both the majority baseline and R-GCN, while $\mathbf{S}^{\mathrm{P}}_{\mathrm{one}}$ remains at baseline level. On *Position*, QS also reaches the top performance (72.2), slightly ahead of One (71.1), with both clearly surpassing all baselines. The only exception is *Direction*, where QS slightly degrades on Mixed-LLM (48.9 vs. 50.0), and both variants remain below R-GCN, although still competitive with other baselines. Comparing datasets, Mixed-LLM yields consistent improve-

ments for Position and Deviation, but degrades performance for Direction with QS. This suggests that higher annotation agreement often stabilises or improves performance, but its impact remains task-dependent and may still reduce performance. Improving annotation quality may therefore reinforce some evaluation patterns while weakening others.

Table 6 shows that, unlike the non-parametric setting, operator distributions become much sharper, with clear dominance within each family. The three tasks are now more clearly separated, and *Deviation* and *Direction* no longer share similar configurations. This increased selectivity is partly due to the smaller number of candidate configurations (except for Deviation in One) and to learning, which promotes more discriminative operator subsets. However, dominant operators are not necessarily consistent between Single-LLM and Mixed-LLM, suggesting that learning remains data-dependent.

Operator	Deviation		Direction		Position	
	One	QS	One	QS	One	QS
ι_{pmax}	0.21 / 0.25	**1.00 / 1.00**	0.00 / 0.00	**1.00** / 0.00	0.25 / 0.33	0.08 / **1.00**
ι_{lin}	0.16 / 0.18	0.00 / 0.00	0.00 / **1.00**	0.00 / **1.00**	**0.75 / 0.67**	**0.85** / 0.00
ι_{euler}	**0.63 / 0.58**	0.00 / 0.00	**1.00** / 0.00	0.00 / 0.00	0.00 / 0.00	0.08 / 0.00
$\alpha_{\text{w}\sum}$	0.38 / 0.15	**0.75 / 1.00**	0.00 / 0.00	0.00 / 0.00	**1.00 / 0.67**	**0.92** / 0.00
$\alpha_{\text{w}\prod}$	0.00 / **0.69**	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
α_{top}	**0.62** / 0.16	0.25 / 0.00	**1.00 / 1.00**	**1.00 / 1.00**	0.00 / 0.33	0.08 / **1.00**
ρ_{pow}	**0.42 / 0.54**	**0.75 / 1.00**	0.00 / 0.00	0.00 / 0.00	0.50 / 0.33	0.08 / 0.00
ρ_{wpsum}	0.28 / 0.24	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.50 / 0.33	**0.77** / 0.00
ρ_{min}	0.30 / 0.22	0.25 / 0.00	**1.00 / 1.00**	**1.00 / 1.00**	0.00 / 0.33	0.15 / **1.00**
$\odot^q_{\text{prod}}$	**0.63** / 0.33	**1.00 / 1.00**	0.00 / 0.50	0.00 / **1.00**	0.25 / 0.33	0.15 / **1.00**
$\odot^q_{\text{wavg}}$	0.21 / 0.33	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.25 / 0.00	**0.85** / 0.00
$\odot^q_{\text{wmax}}$	0.16 / **0.34**	0.00 / 0.00	**1.00** / 0.50	**1.00** / 0.00	**0.50 / 0.67**	0.00 / 0.00
$\odot^s_{\text{prod}}$	–	0.25 / **1.00**	–	0.00 / **1.00**	–	0.31 / 0.00
$\odot^s_{\text{wavg}}$	–	0.25 / 0.00	–	0.00 / 0.00	–	**0.38** / 0.00
$\odot^s_{\text{wmax}}$	–	**0.50** / 0.00	–	**1.00** / 0.00	–	0.31 / **1.00**

Table 6: Top-ranked parametric operator frequencies (%). Values are reported as Single-LLM / Mixed-LLM. Bold indicates the highest value when it is unique for a given operator group, task, and scheme; underlined values denote highest values in case of ties. Number of combinations: Single-LLM — Deviation (One: 120, QS: 4), Direction (One: 1, QS: 1), Position (One: 4, QS: 13); Mixed-LLM — Deviation (One: 118, QS: 1), Direction (One: 2, QS: 1), Position (One: 3, QS: 1).

5.2 Robustness and Best Configurations

To assess robustness to optimisation variability, we retrain selected learnable semantics under multiple random initialisations. In the main experiments, semantics were initialised as their non-parametric counterparts (i.e., all weights set to 1); here, we retrain them over 10 random seeds. We focus on the top-performing configurations identified above: the unique best models for *Position* (72.2, Mixed-LLM, QS) and *Deviation* (48.9, Mixed-LLM, QS), as well as the four configurations scoring 50.0 on *Direction*, including two Mixed-LLM variants (both One) and two Single-LLM (One and QS).

Across seeds, variability remains moderate and task-dependent. *Deviation* exhibits the highest variance ($39.3\% \pm 10.0$, max = 46.7%), reflecting class imbalance and rare large shifts. *Position* remains stable ($65.0\% \pm 4.0$, max = 71.1%), while *Direction* shows moderate variability across configurations, with the following statistics: One (Mixed-LLM) ($40.0\% \pm 6.0$, max = 45.6%), One (Mixed-LLM) ($41.4\% \pm 4.3$, max = 47.8%), One (Single-LLM) ($43.2\% \pm 5.0$, max = 53.3%), and QS (Single-LLM) ($40.7\% \pm 6.2$, max = 54.4%).

Notably, under the best random seeds, *Position* and *Deviation* both decrease slightly (by about one point), whereas *Direction* improves substantially (+4.4, from 50.0 to 54.4).

The strongest configurations are:

- **Deviation: 48.9% (Mixed-LLM, QS, init**$= 1$**)** $\mathbf{E}^{\text{qs}}_{\text{dev}} = \langle \iota_{\text{pmax}}, \alpha_{\text{w}\sum}, \rho_{\text{pow}}, (\odot^q_{\text{psum}}, \odot^s_{\text{psum}}) \rangle$.
- **Position: 72.2% (Mixed-LLM, QS, init**$= 1$**)** $\mathbf{E}^{\text{qs}}_{\text{pos}} = \langle \iota_{\text{pmax}}, \alpha_{\text{top}}, \rho_{\text{min}}, (\odot^q_{\text{psum}}, \odot^s_{\text{max}}) \rangle$.
- **Direction: 54.4% (Single-LLM, QS, best seed)** $\mathbf{E}^{\text{qs}}_{\text{dir}} = \langle \iota_{\text{pmax}}, \alpha_{\text{top}}, \rho_{\text{min}}, (\odot^q_{\text{max}}, \odot^s_{\text{avg}}) \rangle$.

Results on learned gradual semantics focus on operator-level analysis; detailed configuration results, including loss functions and class-weighting schemes, are available at https://github.com/N1N0-P/Learnable-Gradual-Semantics.git.

Statistical significance. We assess the statistical significance of our best gradual semantics against the majority baseline using McNemar's test (McNemar 1947). For *Deviation*, the improvement to 48.9% is not statistically significant ($p = 0.625$), indicating behaviour close to the majority baseline. For *Position*, the improvement to 72.2% approaches the conventional significance threshold of $p < 0.05$, with a measured value of $p = 0.064$, suggesting a strong and consistent gain. For *Direction*, the best result (54.4%, Single-LLM) is not significant ($p = 0.250$). However, these results should be interpreted cautiously due to the limited test size (90 instances), which reduces the statistical power of McNemar's test and makes it difficult to detect significance unless effects are large. Overall, while learnable semantics yield observable improvements—especially for *Position*—statistical significance remains challenging to establish on this dataset.

5.3 Ablation Study: One-Attribute Semantics

To assess the contribution of the multi-attribute representation, we conduct an ablation study in which gradual semantics are restricted to a single node–relation attribute. We focus on the ONE scheme, as QS requires at least two attributes. We perform this evaluation on all 70 possible node–relation attribute pairs (10 node attributes $\times$ 7 relation attributes), using only the best-performing configurations identified in Tables 5 (for Non-Parametric) and 6 (for Parametric).

Task	Single-LLM				Mixed-LLM			
	One-Attr.		Multi		One-Attr.		Multi	
	NP	P	NP	P	NP	P	NP	P
Position	**71.1**	67.8	63.3	**70.0**	66.7	**68.9**	64.4	**72.2**
Direction	**51.1**	42.2	46.7	**50.0**	**50.0**	43.3	46.7	**50.0**
Deviation	**48.9**	47.8	46.7	**47.8**	**47.8**	45.6	46.7	**48.9**

Table 7: Ablation study comparing one-attribute and multi-attribute semantics in non-parametric (NP) and parametric (P) settings.

Analysis (Table 7). In the non-parametric setting, one-attribute configurations outperform the multi-attribute setting across tasks and datasets, showing that individual node–relation attributes can carry strong predictive signal.

This trend does not hold once learning is introduced. In the parametric setting, one-attribute models consistently underperform their multi-attribute counterparts, and their performance almost systematically decreases compared to their non-parametric version (e.g., 51.1 $\rightarrow$ 42.2 on Single-LLM, 50.0 $\rightarrow$ 43.3 on Mixed-LLM for *Direction*). In contrast, multi-attribute models exhibit a non-decreasing behaviour under learning, consistently improving or maintaining performance across all tasks and datasets.

We also observe that one-attribute models are more sensitive to annotation regimes (e.g., Position drops from 71.1 to 66.7 between Single-LLM and Mixed-LLM), while multi-attribute models remain stable (63.3 to 64.4).

Importantly, strong one-attribute performance does not rely on a single dominant node–relation pair. Instead, multiple distinct pairs achieve similar results, and these vary across tasks and datasets. This indicates that the predictive signal is both task-dependent and dataset-dependent, with no pair consistently generalising.

Overall, these results highlight a clear trade-off: one-attribute representations can be strong but are fragile, unstable under learning, and sensitive to annotation choices. In contrast, multi-attribute representations provide a more robust and learnable space, enabling consistent improvements across settings. This leads to two key insights: (i) one-attribute representations can achieve high performance but lack stability and robustness, especially under learning and across annotation regimes; and (ii) multi-attribute representations offer a more reliable and learnable structure, supporting consistent gains across tasks and datasets. Crucially, the existence of multiple equally strong one-attribute pairs indicates that the predictive signal is distributed across attributes rather than concentrated in a single one. This motivates future work on identifying task-adaptive subsets of attributes, bridging the gap between single-attribute and full multi-attribute representations in order to retain signal while reducing noise.

6 Conclusion

We introduced learnable multi-attribute gradual semantics that bridge argument mining and formal reasoning through ADU-level representations enriched with qualitative and strategic attributes. By making evaluation operators trainable, the framework enables data-driven optimisation while preserving the transparency and structure of formal argumentation semantics. Experiments on a new dataset of 600 debates, with two variants (Single-LLM and Mixed-LLM), show competitive performance, sometimes surpassing R-GCN, MLP, LLM-based, and human baselines, while preserving interpretability through transparent evaluation method.

Performance may be influenced by dataset size, class imbalance (notably for Deviation), and noise arising from mismatches between participants' initial stances and assigned debate roles. Future work includes improving relation extraction to assess its impact on persuasion prediction, exploring other informative attributes, extending the operator space toward richer multi-attribute aggregation (e.g. Choquet integrals), analysing the theoretical properties of learned semantics, as well as conducting user studies to assess the impact of the interpretability of our framework on downstream tasks.

AI Declaration

The authors used generative AI tools solely for language assistance. All scientific contributions, analyses, and conclusions are the responsibility of the authors.

Acknowledgements

The work by Victor David was supported by the French government, managed by the Agence Nationale de la Recherche under the Plan d'Investissement France 2030, as part of the Initiative d'Excellence d'Université Côte d'Azur under the reference ANR-15-IDEX-01. Anthony Hunter is grateful for travel funding from the Université Côte d'Azur to facilitate collaboration on this research.

References

Al Anaissy, C.; Suntwal, S.; Surdeanu, M.; and Vesic, S. 2024. On learning bipolar gradual argumentation semantics with neural networks. In *16th International Conference on Agents and Artificial Intelligence*, 493–499.

Amgoud, L., and Doder, D. 2019. Gradual semantics accounting for varied-strength attacks. In *18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019)*, 1270–1278.

Cabrio, E.; Tonelli, S.; and Villata, S. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *International Workshop on Computational Logic in Multi-Agent Systems*, 1–17. Springer.

Cayrol, C., and Lagasquie-Schiex, M. 2005. Graduality in argumentation. *Journal of Artificial Intelligence Research* 23:245–297.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14930–14939.

Habernal, I., and Gurevych, I. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1214–1223.

Hidey, C.; Musi, E.; Hwang, A.; Muresan, S.; and McKeown, K. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*.

Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Yu, P. S. 2020. A survey on knowledge graphs: Representation, acquisition and applications. *CoRR* abs/2002.00388.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Mirzakhmedova, N.; Kiesel, J.; Al Khatib, K.; and Stein, B. 2023. Unveiling the power of argument arrangement in online persuasive discussions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Nabhani, S.; Al Khatib, K.; Pianzola, F.; and Nissim, M. 2025. Storytelling in argumentative discussions: Exploring the use of narratives in changemyview. In *Proceedings of the 12th Argument mining Workshop*, 217–227.

Potyka, N. 2019. Extending modular semantics for bipolar weighted argumentation. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, 273–276. Springer.

Salvi, F.; Horta Ribeiro, M.; Gallotti, R.; and West, R. 2025. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour* 1–9.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, 593–607. Springer.

Stab, C., and Gurevych, I. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 46–56.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 613–624.