CIARAM: Class Imbalance Aware Generative Framework for Relational Argument Mining

Nilmadhab Das, Sayan Pal, V. Vijaya Saradhi, Ashish Anand
Applied Machine Learning (AMaL) Lab, Department of Computer Science and Engineering
Indian Institute of Technology, Guwahati, Assam, India
nilmadhabdas@iitg.ac.in, p.sayan@iitg.ac.in, saradhi@iitg.ac.in, anand.ashish@iitg.ac.in

Abstract

Relational Argument Mining (RAM) is a key task of computational argumentation, which aims to classify the relationships such as *Support* or *Attack* between argument component (AC) pairs. Traditional approaches primarily rely on graph-based modelling with external knowledge sources, which are complex in nature. Also, these approaches struggle with RAM datasets when relation classes are imbalanced, as they are not designed for class-imbalanced scenarios. In this work, we propose ***CIARAM*** framework to reformulate RAM as a text-to-text generation problem to generate relational labels in a flattened text format. To address the class imbalance, we employ a data augmentation strategy using a decoder-only Large Language Model (LLM) to balance the underrepresented relation classes. Across five standard RAM benchmarks, CIARAM produces strong results, specifically with the billion-parameter model, with a substantial gain in performance compared to the latest baseline, demonstrating the strong potential of our approach.

Keywords: Relation Argument Mining, Data Augmentation, Text-to-Text Generation, Class Imbalance

1. Introduction

Relational Argument Mining (RAM) is a specialised task within computational argumentation that focuses on identifying the relationships between pairs of arguments, as shown in Fig. 1. Specifically, given two arguments, the goal is to determine whether $Arg2$ ***Supports*** $Arg1$ or $Arg2$ ***Attacks*** $Arg1$. Unlike traditional Argument Mining tasks, which primarily extract argumentative components and relations (Lawrence and Reed, 2019), RAM seeks to understand the interplay between arguments. RAM has various potential applications, including online debate (Slonim et al., 2021), legal document interpretation (Habernal et al., 2023), opinion aggregation (Cocarascu and Toni, 2017), scientific literature analysis (Fergadis et al., 2021), etc.

The primary challenge of RAM is that the relationship between the arguments is often implicit (Saadat-Yazdi et al., 2023), requiring contextual inference. The diversity in linguistic expressions and domain dependency makes generalization difficult (Cabrio and Villata, 2018). Recently Sun et al. (2022) handles these complexities using graph-based approaches with fine-grained phrase-level similarities (similar words/phrases). Though effective, it overlooks the whole argument-level interaction, where multiple phrase-level interactions are present. More recently, Saadat-Yazdi et al. (2023) uses the culture-specific (domain-dependent) knowledge from external sources to model the discourse dynamics. However, this external knowledge might not be useful for out-of-distribution data where the culture-specific constraints are different. As a result, they often exhibit sub-optimal performance in RAM tasks.

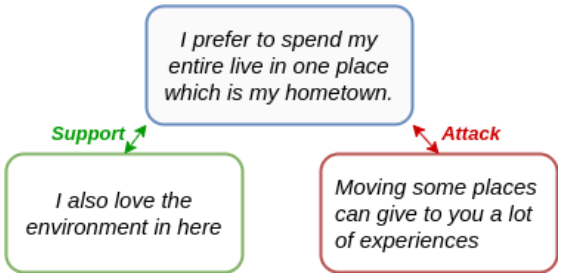

Figure 1: Examples of related argument component pairs taken from Student Essay corpus (Opitz and Frank, 2019) highlighting the *Support* and *Attack* relations.

Another notable challenge in RAM is the presence of class imbalance in widely used datasets (Henning et al., 2023). In many cases, certain argumentative relations are heavily overrepresented, while others occur only sparsely. This skewed distribution can lead models to favour the majority classes, resulting in biased predictions and poor generalisation for the minority classes. Addressing class imbalance is, therefore, crucial to building robust and fair RAM systems. Yet, this challenge remains unaddressed in existing RAM frameworks, leaving a critical gap in the field. Recent advances in large language models (LLMs) have opened up new possibilities for creating synthetic data tailored to specific tasks, making them particularly useful in low-resource or imbalanced settings (Sahu et al., 2022). Compared to traditional augmentation methods, LLM-based approaches have been shown to work better for handling class imbalance (Saad et al., 2025). They can produce

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8096–8105
11–16 May 2026.

contextually accurate and semantically rich examples that closely align with the existing data distributions (Ding et al., 2024). Taking motivation from these works, we apply LLM-based augmentation for class-imbalanced RAM datasets to balance the minority relation classes to improve overall model performance.

With the rise of the generative paradigm, several NLP tasks have been reformulated as text-to-text generation problems, where the input is given as plain text, and the expected output is structured with a flattened representation of target labels. For example, Athiwaratkun et al. (2020a) solved NER and intent classification problems in a unified target sequence. Specifically, *"((AddToPlaylist)) Add [Kent James | artist] to the [Disney | playlist] soundtrack."* is the target sequence of the original input text *"Add Kent James to the Disney soundtrack."*, where the intent is *"AddToPlaylist"* and the named entities are *"Kent James"* and *"Disney"* of type *"artist"* and *"playlist"* respectively. Similarly, Cabot and Navigli (2021) also applied similar flattened output representation to solve relation extraction and classification, producing promising results. Such methodologies have the inherent capabilities to capture the implicit discourse dynamics as well (Oka and Hirao, 2023). A similar methodology is applied by Kawarada et al. (2024) to solve traditional argument mining tasks such as argument component classification and relation classification, which showed improved performances. However, to the best of our knowledge, no existing work has explored the efficiency of RAM tasks within the generative paradigm. Also, the potential of such flattened representations in solving RAM tasks remains unexplored. This gap creates an interesting avenue to investigate their applicability in generative RAM settings.

In this paper, we propose ***C**lass **I**mbalance **A**ware **R**elational **A**rgument **M**ining*, i.e., **CIARAM**, a simple, yet effective text-to-text generation framework for RAM. The input and output of CIARAM is based on the flattened text representation. It also takes care of the minority classes of the class-imbalanced datasets through data augmentation using a Large Language Model (LLM). For the class-imbalanced datasets, we take the instances of the majority class and, using LLM, we perform data augmentation to balance the minority class with the same count as the majority class for that dataset. Thus, CIARAM has three steps: (i) Balancing the minority classes with data augmentation strategy for the class imbalance datasets; (ii) Preparation of flattened representations for both input and output sequences; and (iii) Fine-tuning an encoder-decoder model for the proposed text-to-text generation task with the flattened sequences.

Upon experimentation on five standard diverse-domain RAM datasets including the class-imbalanced ones, CIARAM produces strong results with the billion-parameter model outperforming the existing baselines. In summary:

1. We propose a simple yet effective framework for RAM called **CIARAM**. It is based on the text-to-text generation paradigm, where both input and output are represented as flattened sequences.

2. With the proven ability of LLMs to generate synthetic data for imbalanced settings, we mitigate the class imbalance of several RAM datasets to improve the CIARAM performance, and further release the synthetically generated data with 10% manually evaluated samples to facilitate future research in this direction.

3. CIARAM demonstrates strong performance on five standard datasets , particularly when using the billion-parameter model. An ablation study further highlights the benefits of data augmentation, showing that models with larger parameter counts are more effective at handling class imbalance.

2. Related Work

2.1. Relational Argument Mining

Early works on RAM relied heavily on hand-crafted linguistic features, discourse markers, and syntactic structures (Stab and Gurevych, 2014; Peldszus and Stede, 2015; Stab and Gurevych, 2017). These methods primarily used traditional machine learning classifiers, such as Support Vector Machines (SVMs) and Naïve Bayes classifiers (Palau and Moens, 2009). However, they suffered from scalability issues and limited generalization across domains. With the advent of deep learning, transformer-based architectures such as BERT and RoBERTa demonstrated superior performance by learning contextual representations of argument pairs (Ruiz-Dolz et al., 2021). Multi-task learning frameworks further enhanced Argumentative Relation Classification task (ARC) performance by jointly solving multiple argumentative tasks (Tran and Litman, 2021; Liu et al., 2023). Additionally, models incorporating external commonsense knowledge, such as ARK (Paul et al., 2020) and KE-RoBERTa (Saadat-Yazdi et al., 2023), showed improvements by leveraging knowledge graphs like ConceptNet and WordNet. However, these approaches depend on predefined knowledge graphs, limiting their ability to generalize to unseen arguments. Several recent advancements introduced methods to infer implicit argument relations. For instance, COMET-based approaches generate commonsense inference chains to uncover implicit links

between argumentative units (Saadat-Yazdi et al., 2023). While effective, these methods are constrained by the quality and coverage of pre-existing commonsense knowledge bases. Graph-based neural networks, such as DPGNN (Sun et al., 2022), have also been proposed to incorporate structural dependencies from pre-trained language models (PLMs). These methods improve reasoning but still struggle with capturing nuanced argumentative structures. Another recent method, DISARM (Contalbo et al., 2024), enhances relational argument classification using adversarial training and discourse marker detection. However, it lacks cross-domain evaluation and exhibits instability on smaller datasets, limiting its robustness. Despite these advances, the existing literature lacks RAM formulations that leverage the generative capabilities of LLMs. In particular, no prior work has explored how LLMs can be employed to model RAM tasks, or how their synthetic data generation abilities can be harnessed to address persistent challenges such as class imbalance. We are the first to integrate LLM-based generative modelling with synthetic data augmentation in RAM to explore this research direction.

2.2. LLM-based Data Augmentation

With the rapid progress of LLMs, recent NLP research has increasingly investigated their potential for generating synthetic data to address data scarcity. Several prior works focus on *task-targeted augmentation*, where pretrained LMs are prompted or fine-tuned to produce class-conditional variations of existing samples for low-resource text classification tasks. For example, Kumar et al. (2020) showed that conditioning transformer-based models by prepending class labels can produce diverse, label-preserving synthetic samples, enabling seq2seq models to outperform traditional augmentation in low-resource scenarios. Sahu et al. (2022) demonstrated that prompting GPT-3 without fine-tuning can generate useful labelled data that boosts classifier performance in few-shot settings. Chung et al. (2023) found that diversity-focused generation strategies combined with human-in-the-loop corrections can yield higher-quality augmented datasets, leading to improved model accuracy over few-shot baselines. Given these developments, LLM-based synthetic data generation can offer a compelling solution for RAM tasks, particularly for addressing class imbalance.

2.3. Flattened Sequence Generation

Recent advances in generative approaches have led to the adoption of flattened text representations for tackling a variety of NLP problems. Athiwaratkun et al. (2020b) demonstrated that such a representation could be used to jointly perform tasks like NER, slot filling, and intent detection within a single generation sequence. Similarly, Zhang et al. (2021) applied this technique to aspect-based sentiment analysis, designing both extraction-style and annotation-style flattened formats. Building on these ideas, Paolini et al. (2021) proposed TANL, a general framework for structured prediction that recasts tasks as text-to-text translation. Kawarada et al. (2024) were the first to adopt TANL for jointly solving ACC and ARC tasks through flattened sequence generation. Despite these advances, the potential of treating RAM as a flattened text generation problem remains unexplored, leaving open opportunities to explore this paradigm for capturing complex argumentative relations.

3. Methodology

Our methodology consists of three key steps as shown in Fig. 2: (i) Data augmentation to handle minority classes of class-imbalanced datasets, (ii) Preparation of flattened representations for both input & output, and (iii) Fine-tuning an encoder-decoder model for the proposed text-to-text generation task.

3.1. Data Augmentation

To address class imbalance in RAM datasets, we use *Llama-3.1-instruct* to generate additional instances for underrepresented relation classes. Given $Arg1$ and $Arg2$ from the majority class, we prompt the model to generate an opposing argument of $Arg2$, to which we call it $Arg3$. As a result, a new minority-class relation is created between $Arg1$ and $Arg3$, holding the exact opposite relation of $Arg1$ and $Arg2$. This process continues until class distribution is balanced. Example instances are shown in Step 1 of Fig. 2. Notably, only the *Support* and *Attack* classes are augmented in imbalanced datasets, while the *none* class remains unchanged. Different relation classes of the training split of each dataset, including the augmented ones, are shown in Fig. 3.

3.2. Flattened Representation

We propose a structured approach to input and output representations to solve the RAM task. The input is formatted as **[**$Arg1$**][ ][**$Arg2$**]**, while the target output is structured as **[**$Arg1$**][**$Relation$**][**$Arg2$**]**, where **[**$Relation$**]** represents the relationship between the $Arg1$ & $Arg2$. Notably, for augmented examples, $Arg3$ is applicable instead of $Arg2$. An illustrative example is given in Step 2 of Fig. 2. This flattened representation guides the model to avoid generating irrelevant text by explicitly presenting

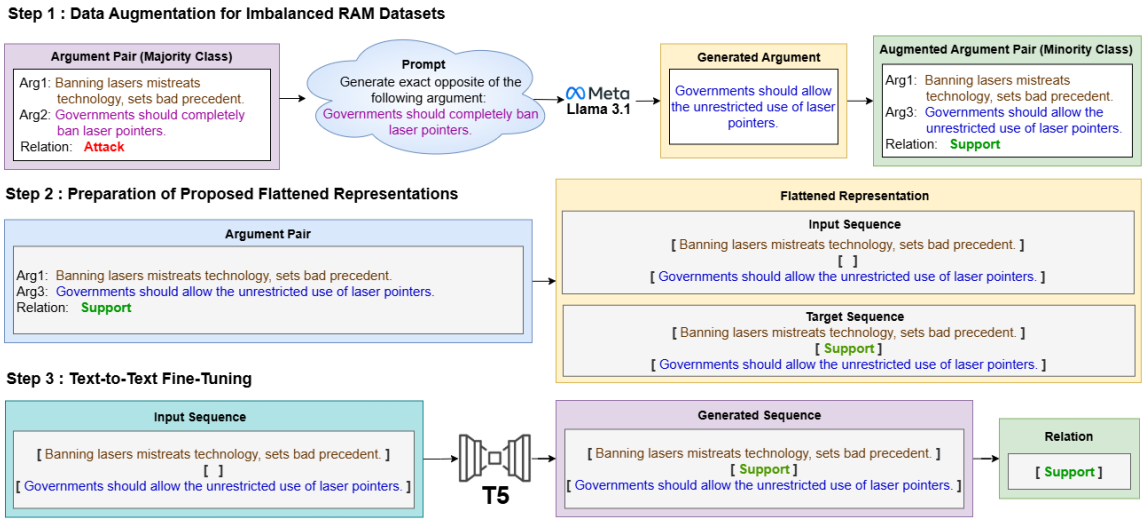

Figure 2: Overview of the proposed **CIARAM** pipeline. **Step 1:** Data augmentation generates counter-arguments for majority-class pairs to synthesize minority-class instances. **Step 2:** Argument pairs are converted into flattened input-output representations with an explicit relation slot. **Step 3:** An encoder-decoder model (T5) is fine-tuned.

Dataset	Train	Dev	Test
Essay	3,070	1,142	1,100
Debate	6,486	2,163	2,162
M-Arg	3,283	410	411
Normative	7,209	1,030	2,060
Causal	5,184	740	1,482

Table 1: Dataset statistics.

both $Arg1$ and $Arg2$ in the input and output sequences. During the generation, the model only needs to fill the empty slot of the input "[ ]" with the relation classes in the output sequence.

3.3. Text-to-Text Fine-Tuning

Using the flattened input sequence, we fine-tune an encoder-decoder model to generate the flattened output sequence as shown in Step 3 of Fig. 2. During the inference, we post-process the flattened output sequence to extract the corresponding relation class of the related arguments.

4. Experimental Setup

4.1. Dataset

We evaluate CIARAM on five publicly available standard RAM datasets as follows:

- **Student Essay (Essay)** (Opitz and Frank, 2019): A corpus of argumentative essays written by second-language speakers, annotated with *attack/support* relations.
- **Debatepedia (Debate)** (Paul et al., 2020): A dataset of structured arguments extracted from Debatepedia, containing pro/con arguments on controversial topics, following a *binary classification scheme (attack/support)*.
- **Presidential Debates (M-Arg)** (Mestre et al., 2021): Transcripts from U.S. presidential debates, annotated with three classes: *support, attack, and none*.
- **Debatepedia-Normative (Normative) and Debatepedia-Causal (Causal)** (Jo et al., 2021): Two subcorpora derived from Debatepedia, containing argument pairs categorized based on normative and causal reasoning. These datasets follow a *binary classification scheme (support/attack)*.

Among these, **M-Arg** and **Essay** exhibit class imbalance for the *Attack* class. Therefore, data augmentation is applied only to these two datasets for the *Attack* class only, while the others remain unchanged.

4.2. Implementation Details

We fine-tune the *Flan-T5* model (Chung et al., 2024) with it's *Base, Large and XL* variants. For *Base* and *Large*, we perform the full fine-tuning with all parameters. For *XL* variant, we use the QLoRA adapter for parameter-efficient fine-tuning. Training was conducted on a single NVIDIA A100 GPU upon five datasets with a learning rate of 0.0005 and a maximum sequence length of 128 tokens. We used a batch size of 64 for both training and

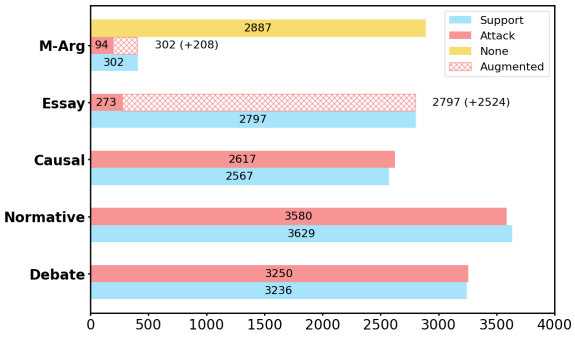

Figure 3: Distribution of different relation classes across the *training* sets of five datasets, including the augmented ones.

Model	Essay	Debate	M-Arg
ARK	60	64	-
KE RoBERTa	70	75	49
RoBERTa+	65.15	74.7	50.37
RoBERTa+ INJ	65.83	74.97	49.35
DISARM (MTL)	69.74	76.14	50.88
DISARM	70.1	76.22	51.34
CIARAM (Base)	49.5	72.4	31.52
CIARAM (Large)	65.28	82.4	39.44
CIARAM (XL)	**83.29**	**89.1**	**58.06**

Table 2: Comparison of Macro-F1 scores of CIARAM with existing baselines. Best scores are in **bold**.

inference, running for 10,000 steps while evaluating every 200 steps to select the best model. Results are averaged over three runs. Following prior studies, we consider **Macro-F1 score** as the evaluation metric for all the experiments. The QLoRA configuration uses a rank (r) of 16 and a LoRA alpha of 32, with a dropout rate of 0.05 and no bias. It is set up for sequence-to-sequence language modeling (SEQ_2_SEQ_LM) and targets specific modules including query, value, key, and output layers. The model loads in 4-bit precision (load_in_4bit) with NF4 quantization (bnb_4bit_quant_type), double quantization enabled (bnb_4bit_use_double_quant), and computes using the torch.bfloat16 data type (bnb_4bit_compute_dtype).

4.3. Baselines

We compare CIARAM with the following SoTA baselines:

- **BiLSTM** (Cocarascu and Toni, 2017): A dual BiLSTM architecture to encode argument component (AC) pairs independently.
- **LSTM-ATT** (Ma et al., 2017): An LSTM with interaction-based attention to enhance AC pair representations.
- **Hybrid-Net** (Chen et al., 2018): A BiLSTM-based model incorporating self- and cross-attention for better argument pair modeling.
- **BERT** (Sun et al., 2022): A vanilla BERT model that uses the `[CLS]` token representation for classification.
- **BERT+LX** (Jo et al., 2021): A BERT-based model that incorporates external linguistic features such as factual consistency and sentiment coherence.
- **BERT+MT** (Jo et al., 2021): A multitask learning-based approach using ARC jointly with textual entailment and sentiment classification.
- **LogBERT** (Jo et al., 2021): A variation of BERT pre-trained on logical reasoning tasks before fine-tuning on ARC.
- **ARK** (Paul et al., 2020): A method that employs a cross-attention mechanism with BiLSTMs and integrates external commonsense knowledge from ConceptNet and WordNet for enhanced argument relation classification.
- **KE-RoBERTa** (Saadat-Yazdi et al., 2023): A knowledge-enhanced RoBERTa model that incorporates commonsense reasoning from external knowledge graphs.
- **DPGNN** (Sun et al., 2022): A dual prior graph neural network that integrates syntactic dependencies and probing knowledge from pre-trained language models (PLMs) for fine-grained argument relation classification.
- **DISARM** (Contalbo et al., 2024): A RoBERTa-based approach that combines multi-task learning and adversarial training by aligning ARC and discourse marker detection (DMD) tasks into a unified latent space. DISARM utilizes the Discovery dataset to learn discourse marker-based representations that improve ARC performance.

5. Results and Discussion

5.1. Main Results

Table 2 and Table 3 compare the performance of CIARAM with its model variants across different datasets against existing baselines. The results show a clear trend: larger model sizes generally lead to better performance on the RAM task. Although CIARAM does not outperform the baselines with *Base* or *Large* variants of Flan-T5 in

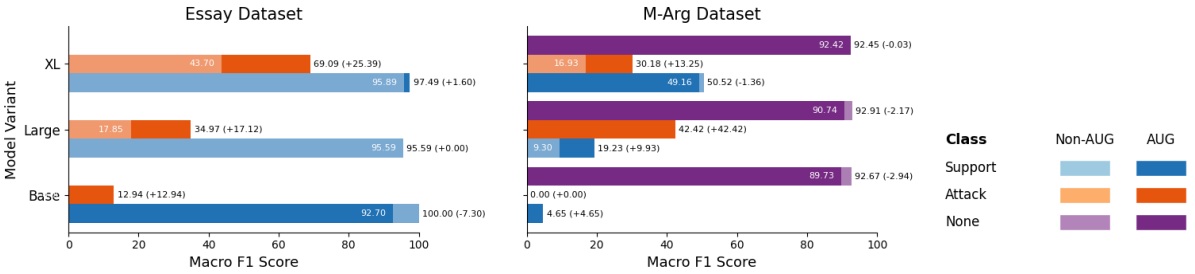

Figure 4: Class-wise Macro-F1 scores for CIARAM across different model sizes (*Base*, *Large*, and *XL*) on the class-imbalanced *Essay* and *M-Arg* datasets, with and without data augmentation. The results highlight the substantial improvements in minority class *Attack*, due to augmentation, particularly as model capacity increases.

Model	Normative	Causal
BiLSTM	71	68.3
LSTM + Att	71.5	70.3
Hybrid Net	67.2	58.8
BERT	79.4	80.7
BERT-LX	78.4	81.5
BERT-MT	79.6	77.5
Log BERT	80.7	80.8
DPGNN	82.9	84.1
CIARAM (Base)	79.4	73.2
CIARAM (Large)	80.7	83.9
CIARAM (XL)	**93.3**	**94.5**

Table 3: Comparison of Macro-F1 scores of CIARAM with existing baselines. Best scores are in **bold**.

most cases, the billion-parameter model *XL* surpasses all the existing baselines in all five datasets. CIARAM *(XL)* achieves macro-F1 improvements ranging from 6.72 to 13.19 points across datasets, highlighting the strength of a flattened text-to-text generation approach over traditional methods. Additionally, the most significant improvement is seen on the class-imbalanced *Essay* dataset, with a gain of 13.19 macro-F1 points. Similarly, on the *M-Arg* dataset, which also suffers from class imbalance, an improvement of 6.72 points is observed. These results underscore the effectiveness of our data augmentation strategy in handling class imbalance, particularly for the minority classes of these datasets.

5.2. Ablation Study

To assess the effect of data augmentation with different model variants of Flan-T5, we compare CIARAM's performance *with* and *without* augmentation on the class-imbalanced datasets: *Essay* and *M-Arg*. Table 4 clearly shows that removing augmented data results in consistent performance drops across all model sizes. The most notable

Method	Model	Essay	M-Arg
CIARAM *(with Aug)*	Base	**49.5**	**31.52**
CIARAM *(w/o Aug)*		49.28 **(-0.22)**	30.84 **(-0.68)**
CIARAM *(with Aug)*	Large	**65.28**	**39.44**
CIARAM *(w/o Aug)*		58.34 **(-6.94)**	32.45 **(-6.99)**
CIARAM *(with Aug)*	XL	**83.29**	**58.06**
CIARAM *(w/o Aug)*		70.04 **(-13.25)**	53.46 **(-4.60)**

Table 4: Ablation study of CIARAM: *with* and *without* data augmentation on class-imbalanced datasets, across model variants. Best scores are in **bold**.

impact is observed with the *XL* variant, where the Macro-F1 score drops by 13.25 points for *Essay* without augmentation. Even with the *Large* model, the effect is evident with significant drops of 6.94 and 6.99 points on *Essay* and *M-Arg*, respectively. The *Base* model experiences relatively smaller but still consistent declines (0.22 and 0.68 points).

In order to further analyse the effect of augmentation upon the minority classes, we additionally compare the classwise F1 scores across all model sizes in Figure 4. With the *Base* model, the *Attack* class is not detected at all in either dataset without augmentation. However, after applying augmentation, the model is able to recognise the *Attack* class in the *Essay* dataset. Moving to the *Large* model, it starts identifying the *Attack* class in the *Essay* dataset even without augmentation. But in the *M-Arg* dataset, the class is still missed unless augmentation is used. Once augmented, the model successfully captures minority classes in both datasets. With the *XL* model, performance improves even further, with the *Attack* class in the *Essay* dataset being handled well both with and without augmentation. Interestingly, in the *M-Arg* dataset, the *Support* class also benefits, and for the first time, a minority class is correctly recognised without any augmentation in this dataset. When augmentation is applied, its F1 score improves even further.

These findings highlight two important observations. First, the effectiveness of data augmentation

Model	Debate	Essay	M-Arg	Normative	Causal
Flan-T5-XL					
Fine-Tuned	**89.1**	**75.15**	**57.26**	**93.3**	**94.5**
Llama-3.1 8B					
Zero-Shot	59.10	45.50	25.00	74.40	69.20
5-shot	77.29	37.92	34.80	71.99	74.79
10-shot	78.83	39.44	36.03	75.28	77.90
20-shot	79.29	42.02	34.15	69.88	82.79
Mistral 7B					
Zero-Shot	49.98	38.04	22.31	55.66	58.32
5-shot	73.77	52.76	42.68	76.72	80.69
10-shot	79.12	50.17	40.69	78.35	83.79
20-shot	73.11	55.28	37.01	78.45	83.73
Qwen 3 8B					
Zero-Shot	47.87	50.00	17.99	29.55	22.03
5-shot	66.22	46.78	25.50	63.51	66.35
10-shot	61.07	41.32	30.34	65.76	70.79
20-shot	61.59	45.10	30.42	64.88	61.28

Table 5: Performance Comparison of the RAM Task: Fine-Tuned *Flan-T5-XL* vs. Zero/Few-Shot *Llama-3.1 8B Instruct*, *Mistral 7B Instruct*, and *Qwen 3 8B Instruct*. Best scores are in **bold**.

becomes more pronounced as the model size increases, indicating that larger models can better utilise the augmented data. Second, augmentation plays a critical role in improving performance on minority classes, particularly in imbalanced datasets, confirming its value in reducing class bias and enhancing overall robustness.

5.3. Zero/Few-shot vs Fine-tuning

To assess the effectiveness of in-context learning for argument mining, we compare the zero/few-shot performance of instruction-tuned models with a fully fine-tuned encoder-decoder model. We choose *Llama-3.1 8B Instruct*, *Mistral 7B Instruct*, and *Qwen 3 8B Instruct* models for the zero/few experiments due to their proven capabilities in zero/few-shot setups. As shown in Table 5, fine-tuned *Flan-T5-XL* consistently outperforms all zero/few-shot instruction-tuned models across every dataset in the RAM task. The most significant gap appears on the Essay dataset, where Flan-T5-XL outperforms the best instruction-tuned model (Mistral 20-Shot) by nearly 20 F1 points, underscoring the limitations of general-purpose LLMs in handling discourse-rich argumentation tasks. Despite gradual gains with more in-context examples, instruction-tuned models fail to close this gap, reinforcing the effectiveness of task-specific fine-tuning for capturing complex argumentative context. While few-shot prompting generally yields modest improvements over zero-shot baselines, several counterintuitive drops occur as the number of in-context examples increases, such as Llama-3.1's decline on Normative from 75.28 F1 (10-shot) to 69.88 F1 (20-shot), Mistral's drop on Debate from 79.12 F1 (10-shot) to 73.11 F1 (20-shot), and Qwen 3's decrease on Causal from 70.79 F1 (10-shot) to 61.28 F1 (20-shot). These degradations mirror the observations of Chen et al. (2023), who report that adding more demonstrations can sometimes reduce accuracy due to prompt interference and spurious correlations.

Model	Augmented By	Supp. F1	Att. F1	Overall F1
$\text{CIARAM}_{\text{Large}}$	LLaMA 3.1 8B	**95.59**	**34.97**	**65.28**
$\text{CIARAM}_{\text{Large}}$	Mistral 7B	92.74	16.00	54.37
$\text{CIARAM}_{\text{Large}}$	Qwen 3 8B	94.71	8.33	51.52

Table 6: Performance of CIARAM using Flan-T5-Large on the *Essay* dataset *with augmentation,* where augmented samples are generated using *Llama-3.1 8B Instruct*, *Mistral 7B Instruct*, and *Qwen 3 8B Instruct*. Best scores are in **bold**.

Dataset	Total	Valid	Percentage (%)
Essay (10%)	341	210	87.0
M-Arg (10%)	24	21	87.5

Table 7: Manual verification of augmented opposite arguments generated using *Llama-3.1-instruct*.

5.4. Analysis and Verification of Augmented Data

In Table 6, we evaluate the effectiveness of data augmentation in CIARAM *(Large)*, where the augmented samples are generated by different LLMs. All configurations use the same methodological flow and differ only in the choice of LLM used for generating synthetic training data. Among the LLMs considered, augmentation using *LLaMA 3.1 Instruct* yields the best results, significantly improving the detection of the minority *Attack* class and achieving the highest overall F1 score. *Mistral 7B Instruct* provides modest performance, while augmentation with *Qwen 3 8B Instruct* leads to a notable decline in both *Attack* and overall F1 score. Based on these outcomes, *LLaMA was selected as the primary augmentation model over Qwen or Mistral* due to its superior quantitative performance along with its consistent generation of linguistically coherent samples.

To further assess the quality of augmented data using *LLaMA 3.1 Instruct*, we manually verified 10% of the generated arguments from both the *Essay* and *M-Arg* datasets. This verification involved evaluating whether each generated argument was contextually valid with respect to the intended relationship type *(Attack)*. As shown in Table 7, 87% of the Essay and 87.5% of the M-Arg augmentations produced by *LLaMA-3.1 8B Instruct* were found to be contextually appropriate. Among the erroneous augmentations, the predominant case was when the generated counterarguments switched topics. This means that when there was supposed to be an Attack relationship between two arguments, the

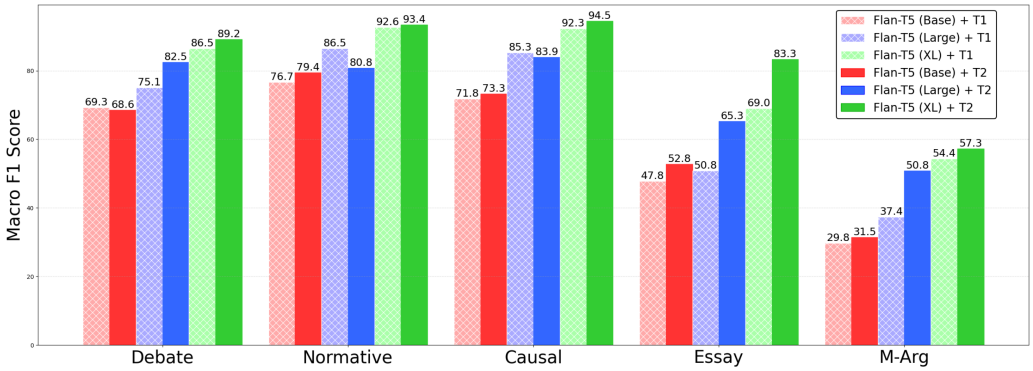

Figure 5: Macro-F1 scores across five datasets using different model sizes with two different output formats: **T1** *(sentence-based, semantically rich)* and **T2** *(originally proposed, concise label-based)*. Surprisingly, the simpler T2 format consistently outperforms the more expressive T1 format across all model sizes and datasets.

new argument was about something completely different. For example, *"Instead of spend time with family and friends, they like to seat on a chair for hours"* was given to create a counterargument, and the generated counterargument was *"Some people do not spend a lot of time watching television"*, which clearly shows the two statements are not focused on the same thing. Another error case was relationship misalignment. This occurs when the generated argument actually *Supports* the given argument rather than *Attacking* it. For example, *"for an engineer or a medicine student I do not find it fair enough to equip gyms instead of their laboratories"* was generated with the given argument *"The university should not provide the best equipment needed for physical education students,"* which actually *Supports* the original argument instead of *Attacking* it. Notably, we did not manually filter these errors, instead we kept them as-it-is with the augmented data. Interestingly, CIARAM achieves strong performance with these slightly noisy (silver standard) augmentations, showcasing robustness in real-world conditions.

5.5. Experiments with Different Input-Output Representation

In order to asses the effectiveness of proposed input-output representations, we conduct experiments with a new representation style. For the output, we use a sentence-like format: *"The Relation between [Arg1] and [Arg2] is [Rel]."* Similarly, the input format is designed as *"The Relation between [Arg1] and [Arg2] is [ ]."* This input-output combination carries richer semantic meaning compared to our originally proposed version. A similar style was used by Paolini et al. (2021), where the authors focused on *relation classification between entities* and got SoTA results in that task. Figure 5 reveals a counterintuitive yet insightful outcome when comparing the performances of these formats: **T1** *(new, sentence-like representation)* and **T2** *(original, concise label-based format)*. Although T1 provides a richer semantic structure by explicitly stating the relation in a natural language sentence, it consistently underperforms compared to the simpler T2 format across all Flan-T5 model variants and datasets, except the *Large* variants in *Normative* and *Causal* datasets. It suggests that while semantically enriched representations like T1 are intuitively appealing, they may introduce unnecessary complexity for models, particularly when the task is to classify relations. The more minimal T2 format allows the model to focus directly on the prediction target, leading to better performance.

6. Conclusion and Future Scope

This paper presents **CIARAM**, a simple yet efficient framework for RAM that leverages the text-to-text generation paradigm, representing both input and output as flattened sequences. To tackle class imbalance in standard RAM datasets, we incorporate a data augmentation strategy using an LLM, which in turn improves the performance of the minority class from the standard RAM datasets. Our experiments with different *Flan-T5* model variants across all five standard datasets showed that the larger models handle RAM tasks more efficiently and captures the class imbalance using the augmented data effectively. One key challenge during data augmentation is the potential for generative models to introduce hallucinations. Sometimes the generated arguments do not accurately reflect the desired outcomes. Manual verification is required to ensure the quality of the augmentation. For the text-to-text generation, we used *Flan-T5* as our backbone model. However, exploring other encoder-decoder models, such as BART, could provide insights into their performance within the current setup.

7. Bibliographical References

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020a. Augmented natural language for generative sequence labeling. In *Proceedings of EMNLP*.

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020b. Augmented Natural Language for Generative Sequence Labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 375–385.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of EMNLP*.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of IJCAI*.

Di Chen, Jiachen Du, Lidong Bing, and Ruifeng Xu. 2018. Hybrid neural attention for agreement/disagreement inference in online debates. In *Proceedings of EMNLP*.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *JMLR*.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Annual Meeting of the Association for Computational Linguistics*.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of EMNLP*.

Michele Luca Contalbo, Francesco Guerra, and Matteo Paganelli. 2024. Argument relation classification through discourse markers and adversarial training. In *Proceedings of EMNLP*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *ArgMining Workshop*.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of EACL*.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *TACL*.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument mining as a text-to-text generation task. In *Proceedings of EACL*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, Suzhou, China. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*.

Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, and Sophia Ananiadou. 2023. Argument mining as a multi-hop generative machine reading comprehension task. In *Findings of EMNLP*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *ArgMining Workshop*.

Yui Oka and Tsutomu Hirao. 2023. Implicit sense-labeled connective recognition as text generation. In *Findings of EMNLP*, Singapore.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *ArgMining Workshop*.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *ArXiv*.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative relation classification with background knowledge. In *Comma*.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of EMNLP*.

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barbera, and Ana Garcia-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*.

Muhammad Saad, Meesum Abbas, Sandesh Kumar, and Abdul Samad. 2025. HU at SemEval-2025 task 9: Leveraging LLM-based data augmentation for class imbalance. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kokciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of EACL*.

Gaurav Sahu, Pau Rodríguez López, Issam Hadj Laradji, Parmida Atighehchian, David Vázquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *NLP4CONVAI*.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen, Lena Dankin, Lilach Edelstein, Liat Ein Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING: Technical Papers*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*.

Yang Sun, Bin Liang, Jianzhu Bao, Min Yang, and Ruifeng Xu. 2022. Probing structural knowledge from pre-trained language model for argumentation relation classification. In *Findings of EMNLP*.

Nhat Tran and Diane Litman. 2021. Multi-task learning in argument mining for persuasive online discussions. In *ArgMining Workshop*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards Generative Aspect-Based Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510.