SENSEI-ASG: A Challenging Dataset for Argument Summary Graph Parsing

Jonathan Clayton[1], **Marco Damonte**[2], **Robert Gaizauskas**[1]
[1] University of Sheffield, Sheffield, United Kingdom
[2] JetBrains
jaclayton2@sheffield.ac.uk

Abstract

We create and make publicly available a novel dataset for the task of Argument Summary Graph Parsing (ASGP), which we call SENSEI-ASG, based on annotating a subset of the SENSEI corpus. Given an argumentative dialogue, such as might be found in a social media exchange, ASGP is the task of creating an Argument Summary Graph, a data structure consisting of nodes containing summaries of arguments in a dialogue and edges representing argumentative relations between them. We find that a previously available ASG dataset, Debatabase-ASG, is not representative of online debates in language use, dialogue length, or graph complexity. In contrast, Debatabase-ASG was created from a curated debate collection, whereas SENSEI-ASG contains examples of spontaneous debates arising in the comment sections of an online newspaper (namely, *The Guardian*). We achieve moderate inter-annotator agreement on the dataset, with a Cohen's kappa of $\kappa = 0.57$, reflecting the inherent challenges in distinguishing argumentative from non-argumentative text. We propose baselines for the new dataset by fine-tuning Llama 3 for the ASGP task using the two ASGP datasets and an additional out-of-domain argument mining dataset, the Argument Annotated Essays Corpus.

Keywords: Argument Mining, Summarisation, Dialogue

1. Introduction

In on-line news and other on-line fora as well, it is common to find user-generated dialogical argumentative content, for instance in newspaper reader comments sections. Such content can contain valuable insights into current issues, but these are often obscured by redundancy, irrelevance and verbosity. A combination of argument mining and summarisation technologies should be able to help make this valuable content more accessible by discarding irrelevancies and redundancy and bringing any underlying argumentative structure to the fore. However, progress in this area is hindered by the lack of suitable datasets to enable research.

In this paper, we introduce a new, publicly available dataset for the task of Argument Summary Graph Parsing (ASGP). This task, introduced in Clayton et al. (2024), is to generate an Argument Summary Graph (ASG), a data structure representing a summarised version of a debate appearing in a dialogical text. Figure 1 shows an example of an ASG from the corpus which we produce in this work, SENSEI-ASG [1]. Such graphs are proposed to be useful structures for users wishing to understand a complex online debate, for example, policymakers. They have advantages over purely textual summaries in making the structure of argumentative relations clear and explicit, which can be important when addressing lengthy online dialogues.

SENSEI-ASG overcomes several limitations of the only currently existing ASG corpus, Debatabase-ASG (Clayton et al., 2024). Unlike the latter, which was produced by annotating a curated online debate collection, SENSEI-ASG contains genuine examples of spontaneously-occurring online dialogue, which are taken from the SENSEI corpus (Barker et al., 2016), in turn sourced from the website of *The Guardian*[2]. In this work, we describe our corpus creation process, and the statistics of the resulting corpus. We also produce an experimental baseline by fine-tuning an LLM to carry out this task end-to-end.

2. Background

The task of Argument Summary Graph Parsing aims to create a compact, graphical summary of argumentative texts. For people interested in social media debates, such as policymakers, these texts are difficult to engage with due to issues such as lengthy comments, redundancy, and off-topic remarks. Graph structures can address this by making the relations between different propositions in a debate explicit, and by combining redundant comments. These issues have been explored in more detail in Clayton (2026, Chapter 1).

The ASGP task (Clayton et al., 2024) is strongly related to two important streams of work in Argument Mining (AM), namely that of Argument Structure Parsing and that of Argument Summarisation.

[1] github.com/acidrobin/SENSEI-ASG

[2] theguardian.com

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8174–8189
11–16 May 2026.

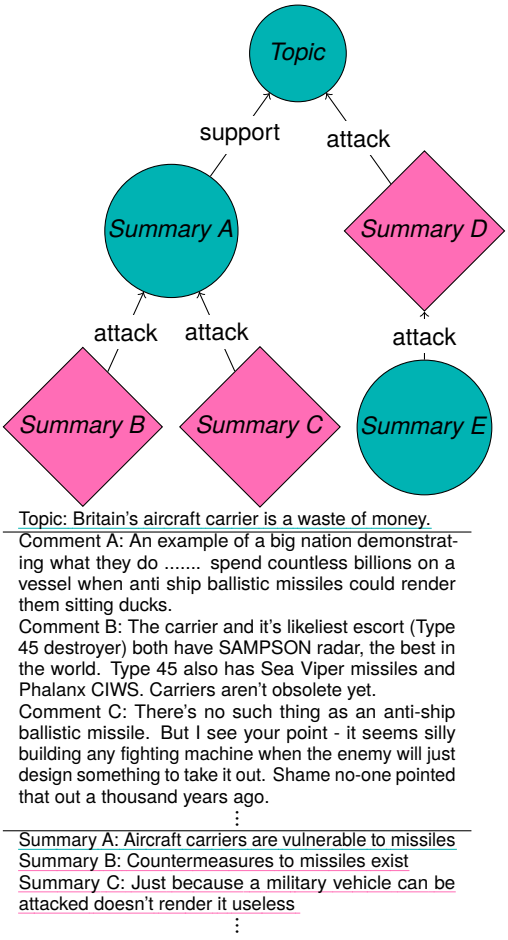

Figure 1: An example Argument Summary Graph from the SENSEI-ASG dataset. Summaries corresponding to each node are shown below the tree for clarity. Summaries in ● support the top-level comment (Topic); summaries in ◆ argue against.

In this section, we will briefly review these two subfields, before describing the ASGP task in more detail, and finally addressing the limitations of the existing ASGP dataset, Debatabase-ASG (Clayton et al., 2024).

2.1. Argument Structure Parsing (ASP)

ASP is the task of extracting structures of reasoning from a text, by identifying the argumentative components (such as, for example, conclusions and premises) that the text contains, and then identifying the relations between them (e.g. support and attack). A wide range of corpora have been developed for this task, including, in the area of dialogue (Park and Cardie, 2018; Ruiz-Dolz et al., 2021; Hautli-Janisz et al., 2022). Numerous models have been designed to tackle this task; here we will only mention Kawarada et al. (2024), who implemented an end-to-end text-to-text approach using an LLM which is similar to the approach taken to the ASGP task in both Clayton et al. (2024) and in this work.

2.2. Dialogic argument summarisation

The summarisation of argumentative dialogues has been addressed from a variety of perspectives; the most conceptually straightforward, which we will call "textual summarisation", involves creating a single textual summary based on a dialogue. Examples of this include Egan et al. (2016) and Irani et al. (2024). These differ from the ASGP task since the summaries generated are not graph-based.

Another important type of summarisation is the task that Altemeyer et al. (2025), in their systematic investigation of the topic, call ArgSum. ArgSum involves, given a text collection containing N arguments, all on the same topic and with the same stance (for example, a collection of "pro" arguments for nuclear power), extracting k argument summaries representing the main arguments in the collection, where $k \ll N$. Altemeyer et al. (2025) further classify ArgSum systems as either clustering-based or classification-based. "Classification based" systems attempt to put arguments into pre-defined topical categories, like Bar-Haim et al. (2020), who match arguments to a list of *a priori* "key points" for each topic (e.g. "Parents are not qualified as teachers" for the topic of "homeschooling"). "Clustering-based" systems, on the other hand, such as Misra et al. (2017) or Li et al. (2024) discover clusters of similar arguments in the data. The ASGP task, however, which involves both generating summaries of comments and identifying the argumentative relations between these summaries, does not fit neatly into either category.

2.3. The ASG Parsing Task (ASGP)

The task of Argument Summary Graph Parsing, introduced in Clayton et al. (2024) is as follows: given an argumentative discussion on topic *t* and containing n comments, generate an ASG with $\mathrm{n}+1$ nodes, where the root is the topic/ Main Claim *M* and the remaining n nodes are summaries of each comment, as shown in the example of Figure 1.

Formally, we can define the ASG parsing task as follows: given a set of comments $C = \{c_1, c_2, \ldots, c_n\}$, where each c_i expresses a stance toward a given main claim M or other comments, the goal is to produce a concise summary s_i for each c_i and construct an argument tree T. The tree T is a directed acyclic graph with M as the root, summaries s_i as nodes, and edges representing support or attack relations between s_i and M, or between summaries s_j and s_i. Optionally, C may also include information about the reply structure of the comments (i.e. C itself may be a tree rather than a list, in which c_j is a child of c_i iff c_j is a reply to c_i in the original discussion).

Note that conversations on social media may contain (i) non-argumentative comments, (ii) irrelevant comments, or (iii) debate on multiple topics. Therefore, in this definition, ASGP would likely form the last step in a pipeline approach after modules that firstly cluster the comments by topic, and secondly remove irrelevant comments.

2.4. Limitations of the Debatabase-ASG Dataset

The only existing ASG dataset to date, Debatabase-ASG (Clayton et al., 2024), has multiple limitations, which we address in this work. The data in Debatabase-ASG is unrealistic, due to the fact that the source it is derived from, Debatabase [3], is a curated debate collection rather than a collection of spontaneous dialogues between real human interlocutors. This means that the resulting data does not reflect real dialogues in several ways:

Formal Language: The language used in the corpus is uniformly written in a clear, formal style, which is likely easier for language models to interpret than more idiomatic texts. Example 1 shows the type of writing which appears throughout the Debatabase-ASG dataset.

> Example 1: Sample comment from Debatabase-ASG, on the topic "This House would make all museums free of charge"
>
> Museums preserve and display our artistic, social, scientific and political heritage. Everyone should have access to such important cultural resources as part of active citizenship, and because of the educational opportunities they offer to people of every age.

In contrast, the language used in actual online arguments is much more complex: looking at the three example comments from the Guardian in Figure 1, we see examples of slang ("sitting ducks"), field-specific jargon ("sea viper", and "Phalanx CIWS") as well as irony (comment C). Each of these features can inhibit the correct processing of a text by an LLM.

Dialogue Length: The dialogues in Debatabase-ASG are extremely short, each being only seven comments long; actual argumentative dialogues, particularly online, can be much longer than this.

Simple ASGs: Due to the short length of the dialogues, and the Debatabase-ASG annotation process, the corresponding Argument Summary Graphs (ASGs) are much less complex than what we might expect from a lengthy dialogue; the maximum tree depth is two, which does not reflect the complexity of graphs found in other subfields of AM (e.g. ASP datasets like Stab and Gurevych 2014, which has an average tree depth of 3.5).

3. Data Source

The SENSEI Annotated Corpus [4] (Barker et al., 2016) is a dataset containing both news articles and online comments from the Guardian news website. In total, it contains 18 news articles, each coupled with approximately 100 comments from the comments sections attached to these articles. They have been annotated by a total of 15 reviewers, with each article being annotated by two reviewers.

We have chosen this existing dataset as a source for our corpus for two main reasons; firstly, it is a good source of naturally occurring debate (albeit, moderated according to *The Guardian*'s Community Standards[5]). Secondly, the corpus contains a few different types of annotations which make the task of generating ASGs from the data more straightforward.

The original purpose of the corpus is for developing tools to create summaries of discussions in online comments. Various types of annotations are available, including textual summaries of the entire comment thread. However, two of the annotation types in the original corpus are useful for our purposes. The first are what the SENSEI annotators called "LABELS", which are per-comment summaries. The second are what they refer to as "GROUPS" and "SUBGROUPS", which are comment-clusters containing topically related comments.

4. Corpus Creation Process

In this section, we describe the process by which we annotated the data. Parts of this pipeline process simply required automatically transforming the existing annotations of SENSEI into a more useful format for our purposes, while other parts required additional human annotation. This was a seven-stage pipeline, which we describe below.

Importantly, we use the comment thread structure to define the argumentative relation structure in the ASG (i.e. which comments attack or support other comments). Doing so saves a significant

[3] idebate.net

[4] Accessible at sensei.sites.sheffield.ac.uk/corpus

[5] theguardian.com/community-standards

GROUP: **Phone hacking**

Label for comment 29: *Police will easily be able to access information*

Label for comment 30: *Only 'normal' people would be vulnerable*

GROUP: **Vulnerable children**

Label for comment 31: *Social workers need a bigger involvement in child care so that police will not have to deal with incidents*

Label for comment 32: *Police having medical records will not help*

Figure 2: Example of the annotations available in the original SENSEI corpus.

amount of annotation time, since for each ASG we only need annotate *n* relations (where *n* is the size of the ASG), instead of n^2.

All manual annotation was done by a single expert annotator. A second expert annotator then annotated a sample of the corpus in order to measure inter-annotator agreement. A Cohen's Kappa of κ = 0.57 was achieved, indicating moderate agreement. This value is commensurate with a number of other prominent Argument Mining corpora, reflecting the fact that these labels have a degree of subjectivity (see Lawrence and Reed 2020, p. 785, a review of AM containing a list of these prominent corpora). We discuss this further below in the description of Step 6, the relation annotation step.

Step 1: Comment Cluster Extraction

The first step is to extract comment clusters from each comment thread. As stated above, the comment threads contain around 100 comments each. Not all of the comments in a thread are necessarily discussing the same topic; for example, in the comments responding to the article entitled "Supercarrier made in Britain hailed as flagship for Better Together campaign", some users debated the utility of the aircraft carrier, while others debated Scottish independence.

As described in Section 3 above, the original SENSEI dataset contains annotations of GROUPS and SUBGROUPS which already roughly cluster the comments sections by topic. We take these clusters as a starting point for our topical comment clusters, using SUBGROUPS where these exist. Note that SENSEI contains annotations from two separate annotators per comment section (from a pool of 15 annotators total); for each comment section we chose one of the two annotators' clusters at random.

However, we note that certain clusters in the dataset are unsuitable for our purposes since they contain only a single "stance" in an argument: for example, some annotators gave clusters names such as "pro-independence". Since we are interested in extracting both stances, we choose to add to every cluster all "descendants" of each original member in the cluster. By "descendants", we mean comments which reply to those comments, the replies to the replies, and so on. This step increases recall for the topics which we are interested in, for example "Scottish independence", since the "anti-independence" comments are often responses to "pro-independence" comments.

Any irrelevant comments which might be included by this step are later filtered out in **Step 6**, a manual annotation step.

Step 2: Remove Redundant Clusters

Due to the fact that we inserted additional comments to the existing clusters in Step 1, some comments may appear in multiple clusters; in other words, clusters overlap. To avoid redundancy in the dataset, we removed clusters which overlapped excessively.

Our procedure involved automatically examining all pairs of clusters. Taking a pair of clusters as two sets of comments, A and B, if $|A \cap B| \geq 0.6 \cdot |A \cup B|$, we removed the smaller of the two clusters from the dataset.

Step 3: Generate an unlabelled argument structure graph

The next step is to generate an unlabelled argument structure graph for each comment cluster. Since, as mentioned, the comments on the Guardian website are tree-structured, with a comment's parent being the comment that it replies to, we can represent each comment cluster as a set of *n* subtrees.

We then combine all *n* subtrees to form a single tree, as shown in Figure 3. We do this by making *t*, the cluster topic, the root of the tree, and attaching the roots of all subtrees to *t*. Note that *t* is simply the topic label acquired from the original SENSEI dataset. Many of these are not suitable for an ASG, since we want the root to be an argumentative proposition which other nodes attack or support. Therefore, in a later step (**Step 5**), we compose our own topic labels.

Step 4: Filter non-Argumentative Clusters

The next step is to identify which of the clusters that we have identified actually contain argumentation. To do this, we used manual annotation. In order to do this, we had to create a definition of an "argumentative" comment for our purposes.

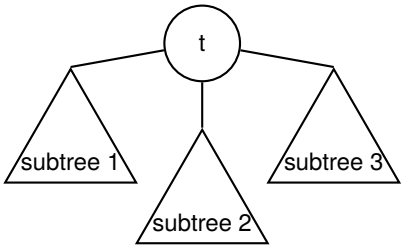

Figure 3: An unlabelled argument structure tree

We define an “argumentative" comment as a comment that does either of the following:

i. Puts across a controversial claim, i.e. a claim that not all participants in the dialogue agree with.

ii. Attacks or supports a controversial claim made by another user.

In order to decide whether the tree as a whole was argumentative, we examined the depth-1 comments, that is, the comments that are children of the root, *t*. If 50% or more of these comments were argumentative, we classified the tree as a whole as “argumentative"; otherwise, as non-argumentative, in which case it is removed.

Step 5: Compose Topic Labels for each cluster

This step involves changing the topic labels from the SENSEI corpus, which are somewhat unhelpful for the ASG task, as they typically just consist of a word or noun phrase describing the general subject of the comments, rather than a specific argumentative proposition that can be argued for or against.

In order to create the topic labels, the annotator took all the level-1 comments from the ASG, and then attempted to formulate a controversial topic, to which the majority (or all if possible) of these comments appeared to be responding.

Depth-1 Comments
Comment 1: There is no need for warships, they are outdated Comment 2: Ships are waste of money - they will need renewing Comment 3: Warships act as an effective deterrent
Argumentative Topic Label:
Building a new warship is a waste of money.

Table 1: Depth-1 Comments and manually annotated topic label from SENSEI

Table 1 shows an example of depth-1 comments and a manually generated argumentative topic label. In this example, it was possible to write a label such that all the depth-1 comments can be interpreted as a relevant response. Where not all the depth-1 comments had an obvious connection, the annotator composed a topic that is relevant to as many as possible. Irrelevant depth-1 comments were subsequently pruned out after being labelled as such in the next step.

Step 6: Annotate Relations

The penultimate step is to label every edge within the output graph. The annotation was done manually using a custom Python script [6].

The annotator used an interface which showed a pair of comments, labelled “Comment A" and “Comment B", as well as two summaries thereof. They were prompted with the question: “Comment B is a reply to Comment A. What is the stance of Comment B towards Comment A? [Attack/ Support/ NA]". [7] For each pair of comments, the child comment must be labelled as either “attacking", “supporting" or "NA" towards its parent. The “NA" category subsumes comments which were either non-argumentative, took a neutral stance toward their parent comment, or were unclear. The full annotation guidelines, which specify how labels were to be decided in different edge cases, are included in *Appendix C. Annotation Guidelines*.

Step 7: Prune Non-Argumentative Comments and Format Data

We now create the final dataset using the annotations carried out in the previous step. Where a child comment was annotated as “NA", the subtree consisting of that comment and all its descendants is pruned from the tree. The rest of the edges are assigned the corresponding “support" or “attack" labels given by the annotator.

We then use the trees to create inputs and outputs for the parsing task formatted as in Table 4. This format is identical to the one in the Debatabase-ASG dataset (Clayton et al., 2024). The input contains only the main topic and the comments, without any of the reply structure. The output contains a summary of each comment coupled with an indication of its parent in the tree, and the relation to the parent, in parentheses.

5. Corpus Statistics

Our final dataset contains 70 comment threads paired with ASGs. This dataset is smaller than Debatabase-ASG in terms of number of ASGs (70 vs 590). However, as shown by the statistics in

[6] This script is provided at github.com/acidrobin/sensei_annotation.

[7] An example of a prompt with comments is shown in Figure 6, *Appendix C. Annotation Guidelines*.

	Corpus	
	SENSEI-ASG	Debatabase-ASG
Size of Dataset	70 threads	590 threads
Train: Val: Test split	57 : 7 : 6	472: 59: 59
Average Tree Depth	5.7	2
Max Tree Depth	12	2
Min Tree Depth	2	2
Average N. Nodes	16.4	7
Avg. Comment/ ADU Length (word tokens)	61.6	176
Avg. Node Summary Length (word tokens)	20.9	17.8
N. Attack relations	797	2477
N. Support relations	302	1864

Table 2: Descriptive statistics comparing the two ASGP corpora used in our experiments.

Table 2, the graphs in SENSEI-ASG are significantly more complex on average than those in Debatabase-ASG, when measured either by average number of nodes in the tree (16.4 vs 7) or by the average depth of these trees (5.7 vs 2). Figure 4 shows the distribution of topical categories in SENSEI-ASG, indicating that a broad range of topics have been included.

		Annotator B		
		Support	**Attack**	**N/A**
Annotator A	**Support**	59	13	7
	Attack	1	116	6
	N/A	32	19	29

Table 3: Confusion matrix between Annotator A (rows) and Annotator B (columns).

5.1. Inter-Annotator Agreement

As discussed above, a second annotator annotated a sample of the labels, in order to validate our annotation guidelines. This sample consisted of 20 comment clusters (out of 70 total), or 282 comment pairs.

A Cohen's Kappa of $\kappa = 0.57$ was achieved, indicating moderate agreement. The two annotators generally agreed on distinguishing Support from Attack labels, with 59 and 116 instances correctly agreed for Support and Attack, respectively. The primary disagreements arose when differentiating between argumentative labels (Support/Attack) and non-argumentative segments (N/A), particularly for segments labelled as N/A by one annotator but as Support by the other (32 instances) or as Attack (19 instances). This may reflect an inherent difficulty in identifying whether a comment is relevant to another, something that often depends on inherently ambiguous pragmatic assumptions. Table 3 shows the full confusion matrix between the two annotators.

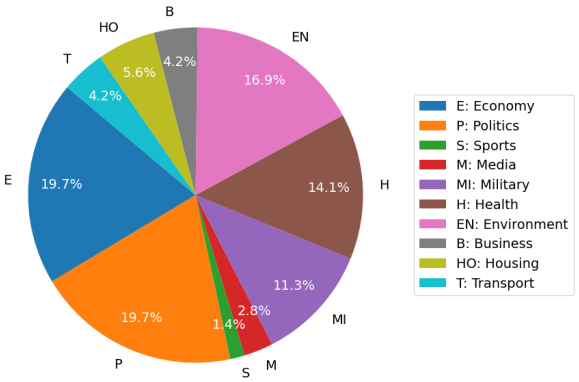

Figure 4: The distribution of topic categories in the SENSEI-ASG dataset

6. Evaluation Metrics for the ASGP Task

Following Clayton et al. (2024), we use three different metrics to evaluate separate aspects of the output produced by a candidate system against the SENSEI-ASG gold standard. However, we propose a new metric to replace Graph Edit Distance for the evaluation of graph structure, which we call Node Position F1. Below, we will describe each of these three metrics in turn.

The three aspects that are evaluated are (1) the correctness of the graph structure, (2) the correctness of the edge labels (i.e. attack/ support), and (3), the quality of the summary text.

Node Position F1: We introduce a novel metric, Node Position F1, to evaluate aspect (1), correctness of the graph structure. In this we differ from Clayton et al. (2024) who evaluated this using the Graph Edit Distance (GED). While GED is a suitable metric, it can be difficult to interpret since it ranges between zero and infinity, with a graph identical to the reference having a score of zero. In contrast, Node Position F1 ranges between 0 and 1, and a graph with an identical structure to the reference has a score of 1.

Node Position F1 scores graphs based on the position of each node relative to the root: nodes are considered to be positioned correctly if the path from the node to the root is identical in the predicted graph and the reference.

As shown in Figure 1, each node in the tree has a label, such as “Comment 1". We define a function $lab(v,T)$ which returns the label of a given node v in a tree T. For each node v, we can then also define the root-to-node path, denoted by $P(v,T)$, as

$$P(v,T) = (\text{lab}(v_0), \text{lab}(v_1), \ldots, \text{lab}(v_k)) \quad (1)$$

where $v_0 = v$, $v_k = \text{root}(T)$, and for each i such that $0 \leq i < k$, v_{i+1} is the parent of v_i in T.

The set of all node-to-root paths $\mathcal{P}(T)$ in T is then the collection of all such sequences for each node $v \in V$:

$$\mathcal{P}(T) = \{P(v,T) \mid v \in T\} \quad (2)$$

We can then define the sets of all node-to-root paths in the ground truth tree T_{true} and the generated tree T_{pred} respectively:

$$P_{true} = \mathcal{P}(T_{true}) \quad (3)$$

$$P_{pred} = \mathcal{P}(T_{pred}) \quad (4)$$

We can then define True Positives (TP), False Positives (FP) and False Negatives as below:

$$TP = |P_{\text{true}} \cap P_{\text{pred}}| \quad (5)$$

$$FP = |P_{\text{pred}} - P_{\text{true}}| \quad (6)$$

$$FN = |P_{\text{true}} - P_{\text{pred}}| \quad (7)$$

Finally, the F1-score (van Rijsbergen, 1979) is calculated as follows:

$$F1 = 2 \cdot \frac{TP}{2TP + FP + FN} \quad (8)$$

Node-stance F1: Node-stance F1, which was introduced in Clayton et al. (2024), indirectly evaluates the correctness of the attack/support relations predicted by the model by evaluating the correctness of the stance of each node towards the root.

ROUGE: The quality of the summary text is evaluated using the ROUGE score (Lin, 2004). ROUGE is a commonly used metric in text summarisation, which works via measuring n-gram overlap. Following Clayton et al. (2024) we use ROUGE-2, and calculate ROUGE scores between the generated and gold standard graph representations as a whole string (as in the right-hand column of Table 4).

7. Baselines

In this section, we describe the Llama 3 models fine-tuned to provide baselines for this dataset.

7.1. Fine-Tuning Datasets

We use three different datasets to fine-tune the model: the two ASG datasets (SENSEI-ASG and Debatabase-ASG) as well as an out-of-domain dataset. For the out-of-domain dataset, we use a popular Argument Structure Parsing dataset, the Argument Annotated Essays Corpus (AAEC) (Stab and Gurevych, 2014). This corpus is unlike the ASGP corpora since the task only involves extracting structure rather than producing a summary.

7.2. Formatting for LLM Training

As in Clayton et al. (2024), we formulate the task as an end-to-end text-to-text task. Given this, the formatting of the inputs and outputs is important. For the two dialogical corpora, Debatabase-ASG and SENSEI-ASG, we use the same formatting as in our 2024 paper. Table 4 illustrates this; as shown in the table, the input consists of the main topic followed by a concatenated list of comments, each with associated comment numbers. The output formatted as a list of summaries. Each summary also comes with an associated comment number, in addition to, in parentheses, the parent of that comment in the ASG, as well as the relation to the parent (“support” or “attack”).

Example 2: Our formatting of the Argument Annotated Essays Corpus

Input: As far as I am concerned, we should do our parts to do what ever it takes to protect old buildings, letting next generations still have a chance to look at them. Here are some reasons and examples to illustrate my viewpoint. Historic buildings have intrinsic values, because they are significant symbols of a city.

Output:
MajorClaim: we should do our parts to do whatever it takes to protect old buildings, letting next generations still have a chance to look at them.
Claim 1 (supports MajorClaim): Historic buildings have intrinsic values.
Premise 1 (supports Claim 1): they are significant symbols of a city.
Premise 2 (supports Claim 1): Old buildings are reminders of a city’s culture and complexity.

We create a similar format for the AAEC dataset, in order to make it easier for models trained on one corpus to be applied to another. Example 2 illustrates this format. There are some small differences from the formatting in Table 4; firstly, the input is an entire essay rather than a list of comments; secondly, since the output is “Argumentative Discourse Units” rather than comments, these have labels such as “Premise 1” or “Claim 2”, instead of a comment number, and finally, each of the ADUs directly corresponds to a text span in the original essay, rather than being a summary as in Debatabase-ASG and SENSEI-ASG.

Example Input	Example Output
Main topic: South-Eastern England gets disproportionately high levels of public funding. Comment 1: The most heavily subsidised rail users in the region getting nearest the targets are given yet more disproportionate subsidy... Comment 2: It's the same with all government funding - be it transport, arts, or whatever. London and its surrounds always get a disproportionately large slice of the pie! Comment 3: Are you sure that per capita London is better funded than other parts of the country? I have my doubts. Comment 4: Yes, just do some searching on the web. There's plenty of information out there! Comment 5: Right, take money from all rail travellers and give it to the SE commuters.	Comment 1 (supports Main topic): SE getting disproportionate public funding; vote buying Comment 2 (supports Comment 1): London getting disproportionate public funding Comment 3 (attacks Comment 2): questions London getting disproportionate public funding Comment 4 (attacks Comment 3): London getting disproportionate public funding – agrees – look on web for article Comment 5 (supports Comment 1): SE getting disproportionate public funding

Table 4: Example inputs and outputs from the SENSEI-ASG Dataset

7.3. Training Details

We carried out fine tuning using the two different test sets, SENSEI-ASG, and Debatabase-ASG, and all seven different possible combinations of three training sets – SENSEI-ASG (S), and Debatabase-ASG (D) and AAEC (A) – which we will denote {S, D, A, S+D, S+A, D+A, S+D+A}. We carried out a separate hyperparameter tuning run using grid search for each training set/ test set pair, using the appropriate validation data for each target test set; hence, we carried out a total of 14 grid search runs.

We chose to train the Llama 3-8B model (Grattafiori et al., 2024) for this task. This is a logical choice due to the high performance of Llama 2 in Clayton et al. (2024); we confirm that Llama 3 performance exceeds that of Llama 2 in experiments on Debatabase-ASG (Section 8).

Fine-tuning was conducted on a high-performance computing (HPC) node containing 8 NVIDIA V100 GPUs. Due to the large size of the Llama 3-8B model, we used QLORA (Dettmers et al., 2023), a parameter-efficient fine-tuning technique. We fine-tuned for 8 epochs, using the AdamW optimiser (Loshchilov and Hutter, 2017). Early stopping was employed with a patience parameter set to 2, meaning training was halted if the validation performance did not improve for two consecutive epochs. The hyperparameters that we tuned were learning rate, from the set $\{1e{-}4, 1e{-}5\}$, and weight decay, from the set $\{1e{-}2, 1e{-}3, 1e{-}4\}$. The other hyperparameters were set to the values shown in *Appendix A. Llama 3 Model Parameters*.

8. Results

Llama 3 Performance on Debatabase-ASG Table 5 compares the best-performing models from Clayton et al. (2024) to the Llama 3 model used in this work, both fine-tuned on Debatabase only, and fine-tuned on all three datasets. In both cases, the Llama 3 model outperformed Llama 2, in all metrics, except for ROUGE for the Debatabase-only model.

Model	ROUGE ↑	S-F1 ↑	GED ↓
Clayton et al. (2024)			
Phi-2	0.5460	0.7825	2.5667
Llama 2	**0.5829**	0.8700	1.45
GPT-3.5	0.5422	0.8333	1.85
This Work			
Llama 3-D	*0.577*	*0.921*	*1.21*
Llama 3-S+D+A	0.535	**0.955**	**0.85**

Table 5: End-to-end (fine-tuned) results on Debatabase-ASG. Higher ROUGE-2 and Stance-F1 are better (↑), lower GED is better (↓). GED is used here instead of to allow comparison with Clayton et al. (2024). The best-performing results are shown in **bold**, and the second-best in *italics*.

Model Performance on Debatabase-ASG vs SENSEI-ASG Figure 6 compares the performance of fine-tuned Llama 3 on two different test sets: Debatabase-ASG and SENSEI-ASG. We compare the model which performed best on each test set, which in both cases happened to be the model fine-tuned on all three datasets.

The comparison shows that SENSEI-ASG is much more challenging than Debatabase-ASG, with the model achieving around half the performance on the former dataset compared to the latter.

Fine-tuning with Additional Datasets Figure 5 shows the performance of Llama 3 on the SENSEI-ASG test set, when trained on various combinations of training sets (a table with exact scores is provided in *Appendix B. Full Results*.

Overall, when we consider those models that performed well across the three metrics, it would appear that the best performing combinations of training sets are S+D, S+A and S+D+A; these are the combinations which contain both in-dataset data (SENSEI) and additional training set(s).

Another interesting observation is that performance does not seem to improve when a third

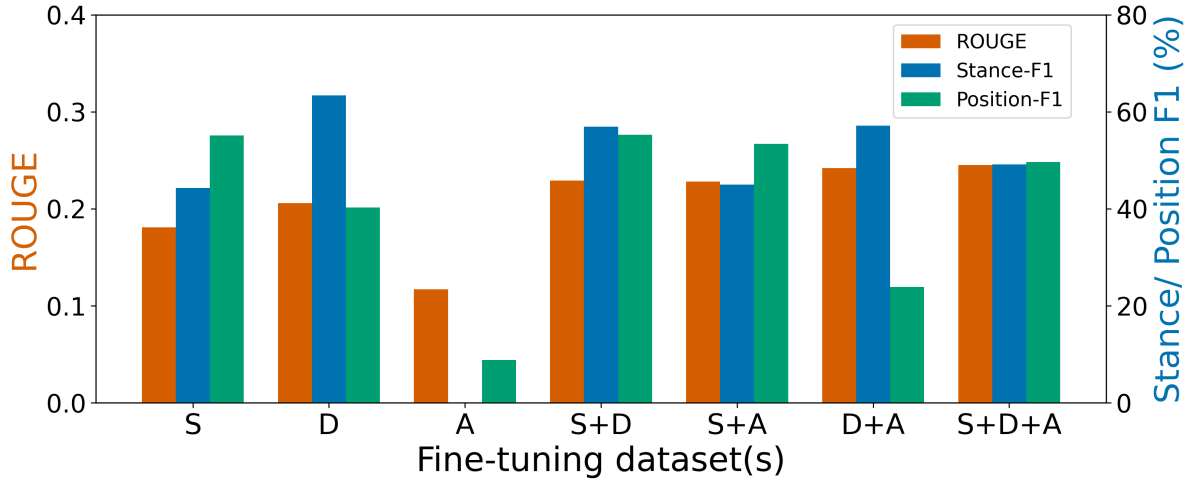

Figure 5: Performance of the fine-tuned Llama 3 model with different fine-tuning sets. Note that the y-axis on the left is for ROUGE and the axis on the right is for stance and position F1

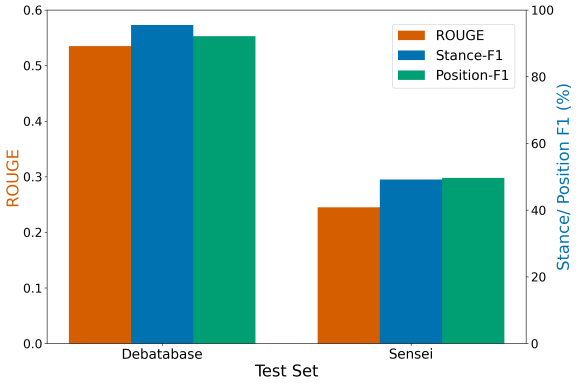

Figure 6: Performance of the best model (Llama 3 S+D+A) compared on the Debatabase-ASG and SENSEI-ASG test sets

dataset is added (S+D+A) compared to either S+D or S+A. This could reveal a limit to performance gains which can be achieved by adding training data that differs qualitatively from SENSEI-ASG.

9. Conclusion

In conclusion, we have created a novel dataset, SENSEI-ASG, for the task of Argument Summary Graph Parsing. This dataset is more realistic than the only existing dataset for this task, Debatabase-ASG, due to the fact that the data contained within it is taken from user comments from the Guardian newspaper instead of a curated online debate collection. This results in several features that make it a more complex and challenging dataset to work with: the writing style is less formal, the comment threads are much longer, and their corresponding argument graphs are much more complex.

We proposed baselines using the Llama 3 model in end-to-end parsing of Argument Summary Graphs. We show that performance scores are much lower on SENSEI-ASG than Debatabase-ASG, reflecting the higher complexity of this dataset. We also found that additional fine-tuning on both Debatabase-ASG and an another dataset for a related task (the AAEC, Stab and Gurevych 2014) can improve performance.

Limitations

- **ASG structure is based on reply structure:** Due to our annotation process, the structure of the ASGs in the output is partly based on (although not strictly determined by) the reply structure in the original comment threads. This may not always reflect the reality of argumentative relations that can be found in a text. Future work should try to build datasets to address this.
- **Annotation Scheme:** Our annotation scheme was designed partly for time and cost-effectiveness, and therefore is extremely simple. A major simplification that we make is annotating at the comment level rather than the sub-comment level.
- **Size:** The corpus is relatively small, consisting of 70 comment threads and their corresponding ASGs. As a result, the experimental baselines reported here may not fully reflect the potential performance of LLMs on this task, which could improve if a larger corpus were developed.

10. Bibliographical References

Moritz Altemeyer, Steffen Eger, Johannes Daxenberger, Tim Altendorf, Philipp Cimiano, and Benjamin Schiller. 2025. Argument summarization and its evaluation in the era of large language models. *arXiv preprint arXiv:2503.00847*.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. *arXiv preprint arXiv:2005.01619*.

Jonathan Clayton. 2026. *Graphical Summarisation of Argumentative Text*. Ph.D. thesis, University of Sheffield. Chapter 1.

Jonathan Clayton, Marco Damonte, and Robert Gaizauskas. 2024. Parsing Graphical Summaries from Argumentative Dialogues. In *Computational Models of Argument*, pages 37–48. IOS Press.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized LLMs. *arXiv:2305.14314*.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2024. Wiba: What is being argued? a comprehensive approach to argument mining. *arXiv preprint arXiv:2405.00828*.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument mining as a text-to-text generation task. In *EACL Proceedings (Volume 1: Long Papers)*, pages 2002–2014.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Hao Li, Yuping Wu, Viktor Schlegel, Riza Batista-Navarro, Tharindu Madusanka, Iqra Zahid, Jiayan Zeng, Xiaochi Wang, Xinran He, Yizhi Li, and Goran Nenadic. 2024. Which side are you on? a multi-task dataset for end-to-end argument summarisation and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 133–150, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. 2017. Using summarization to discover argument facets in online ideological dialog. *arXiv preprint arXiv:1709.00662*.

C. J. van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. Butterworths, London. F-measure defined on p. 114.

11. Language Resource References

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtić, Mark Hepple, and Robert Gaizauskas. 2016. The sensei annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 42–52.

Jonathan Clayton, Marco Damonte, and Robert Gaizauskas. 2024. Parsing graphical summaries from argumentative dialogues. In *Computational Models of Argument*, pages 37–48. IOS Press.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3291–3300. European Language Resources Association (ELRA).

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ramon Ruiz-Dolz, Montserrat Nofre, Mariona Taulé, Stella Heras, and Ana García-Fornes. 2021. Vivesdebate: A new annotated multilingual corpus of argumentation in a debate tournament. *Applied Sciences*, 11(15):7160.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Appendix A. Llama 3 Model Parameters

Parameter	Llama-3
Base model	meta-llama/Llama-3-8B
Optimizer	AdamW
Batch size	2
Max length	7000
Dropout	0.1
AdamW beta1	0.9
AdamW beta2	0.999

Table 6: Training Parameters for Llama-3

Parameter	Llama-3
r	64
lora_alpha	16
lora_dropout	0.1
bias	none
task_type	CAUSAL_LM
target_modules	q_proj, up_proj, o_proj, k_proj, down_proj, gate_proj, v_proj
load_in_4bit	True
bnb_4bit_quant_type	nf4
bnb_4bit_use_double_quant	True
bnb_4bit_compute_dtype	torch.bfloat16

Table 7: QLORA parameters for fine-tuning Llama-3.

Appendix B. Full Results

		Training Set						
		S	D	A	S+D	S+A	D+A	S+D+A
Test Set	S	0.181	0.206	0.117	0.229	0.228	0.242	0.245
	D	0.169	0.577	0.059	0.566	0.157	0.581	0.535
	A	0.296	0.121	0.848	0.240	0.862	0.872	0.853

Table 8: ROUGE scores (LLaMA-3)

		Training Set						
		S	D	A	S+D	S+A	D+A	S+D+A
Test Set	S	44.30	63.40	0.00	56.90	45.00	57.10	49.14
	D	35.86	92.09	0.00	91.48	50.48	94.63	95.49
	A	0.00	0.00	49.72	0.00	50.38	51.28	54.98

Table 9: Node Stance F1 (LLaMA-3)

		Training Set						
		S	D	A	S+D	S+A	D+A	S+D+A
Test Set	S	55.12	40.24	8.82	55.28	53.35	23.87	49.65
	D	56.19	90.48	14.29	93.10	69.29	91.43	92.14
	A	8.30	8.30	43.36	8.30	48.13	46.12	45.93

Table 10: Node Position F1 (LLaMA-3)

Appendix C. Annotation Guidelines

C.1. General Guidelines

In this task, you will be shown pairs of comments from online news comments sections.

Each pair of comments will be labelled “Comment A" and “Comment B". In each case, Comment A and Comment B are written by two different users and Comment B is responding directly to Comment A.

Comment A: Who are "the rich"? To some, a person earning 50k is rich. But if they rent in London and have children they won't feel at all rich. Some CIFers define rich as anyone with more money than them. Comment B: I agree with both of your points. I don't think someone earning 50k is rich, probably below average if living in London.	Comment A: There's something wrong if we have parts of our country that people on £50,000 a year can't afford to live in, I earn a decent salary, so does my partner, we're by no means poor but we couldn't afford to live in London even if we wanted to, it's an incredibly unhealthy situation Comment B: Since when, have those earning double the average wage, suddenly become entitled to the consider themselves 'not rich' or what they can't say 'poor' (because it would be unseemly) . What utter bollocks!
Label: **Support**	Label: **Attack**

Table 11: Uncontroversial examples of support (left) and attack (right) relations between pairs of comments.

Discussion Topic: NHS tax is floated by Liberal Democrats to fill £30bn hole

Comment A Summary: Rich - should pay more tax. Low paid workers constantly fear losing home/ assets etc.
Comment A: Great, let the proles squabble amongst themselves which of their number should consider themselves 'rich' (clue anyone who is 'a couple of paychecks away from the street' as they would say in the US, ain't rich), whilst the oligarchs get on with the job of chiselling Olympic sized swimming pools and full-sized ballrooms underneath their London pads and furnishing their yachts with musical waterfalls and flocks of scented sheep and the like. All hail the 'wealth creators' and God forbid that they should pay a fair share of tax, that way lies disaster.

Comment B Summary: Businesses - should pay triple NI tax - people should pay lower income tax
Comment B: Triple the NI for businesses, and lower the income tax. The NI is one tax multinationals can't avoid paying.
Comment B is a reply to Comment A. What is the stance of Comment B towards Comment A? [Attack/ Support/ NA]

Figure 7: An example of the type of comment pair you will be asked to annotate.

As well as the comments, you will be provided with the topic of the discussion, as well as a short summary of each comment intended to aid your understanding of what each user is trying to convey.

Figure 7 is an example of the prompt that you will be shown, which shows the discussion topic and two comments, along with a summary of each of those comments. You are prompted with three options to select from to describe the stance of Comment B towards Comment A: Attack, Support or NA.

In general, we encourage you to bias your selection towards “Attack" or “Support" and select “NA" (meaning “Non-argumentative"/ “Neutral") only as a last resort.

- Select “Attack" if the stance of Comment B generally disagrees with the stance of Comment A.
- Select “Support" if the stance of Comment B generally agrees with the stance of Comment A.
- Select “NA" if the comment:
  1. is not argumentative: for example, it is an explanation, a joke, or an insult
  2. otherwise does not give an opinion about the comment it is responding to
  3. is incomprehensible in context (you cannot understand how Comment B is supposed to relate to Comment A).

C.2. Further Guidelines

This section contains more specific advice on types of comment pairs which may be more challenging to label. We are interested in the broadest possible definition of argument, so if in doubt it is preferable to label a comment as either “Attack" or “Support" rather than “NA".

Rhetorical Questions: Rhetorical questions (in cases where it is clear that a user is trying to make an argument), count as argumentative. For example, the comment in Table 10, containing a rhetorical question, counts as an Attack:

Example Comments	**Comment A:** Britain, super killing machine for the rich since 1707, now alienating ourselves further from a global world by having out fingers in all the dirty pies. **Comment B:** By "killing machine" you mean stuff like the Royal Navy outlawing slavery and piracy on the high seas throughout the C19th?	Inflation adjusted, yes, but in terms of affordability of basic commodities, housing etc, how does that work out?
Label	**Attack**	**NA**
Reason For Label	We can tell from the context that this is a rhetorical question/ challenge, and not a request for clarification, hence we count it as argumentative (an attack)	Questions asking for clarification on some point are not the same as rhetorical questions, and don’t count as argumentative.

Table 12: Rhetorical Questions

Critical comments/ insults: Comments containing insults or criticisms of another user should only be considered “argumentative" if they act as an argument in the context of the discussion. For example, calling another user “illiterate" could be an argumentative comment if they are attempting to call into question that user’s claim to be an authority on climate science, for example. However, if the insult has no obvious argumentative function but seems to be nothing more than an example of online harassment, then it should be labeled as neutral. Note that this is not the same as requiring that the comment be “civil" - we acknowledge that non-civil comments may still have an argumentative function.

Example	Again you must work in the public sector if you think you get big redundancy payments and anything more than a state pension. So that means JSA if your unemployed before 65. You need to get out of your cosy world and look beyond your parents that you envy so much.	You sound lovely
Label	**Attack**	**NA**
Reason for Label	This comment contains criticism/ insult but also arguments, therefore we mark it as argumentative	A comment containing just a personal attack, and nothing more, should be marked as NA

Table 13: Personal Criticism/ Insults

Comments with Mixed Attack/Support relations Where there are parts of Comment B which disagree with Comment A and parts which agree with it, pick the option which seems to express the overall sentiment of the comment. If this is unclear or the user gives equal weight to attacking and supporting different parts of the previous comment, then select “NA".

Example	**Comment A:** He said it was an example of a big nation demonstrating what they do ..... spend countless billions on a vessel that will at best have no aircraft for at least 6-10 years and when there is enough support vessels to defend this hulking lump. Lets gloss over the anti ship ballistic missiles that could render them sitting ducks. **Comment B:** Agree regarding the time scale for fixed wing aircraft, however I'm not so sure about your statement with regards to anti ship missiles. The carrier and the Type 45 destroyer both have SAMPSON radar, the best in the world. Type 45 also has Sea Viper missiles and Phalanx CIWS. Carriers aren't obsolete yet.	**Comment A:** BAe and HMG have been entirely clear on this, that in the event of a Yes vote that is what the outcome will definitely be. The SNP position that the UK would continue to build them in Rosyth is nonsense as a large government investment is needed in the yard for the T26 programme. **Comment B:** I don't dispute your statement, rUK tax payer will just have to pick up the bill if or as and when they do. However According to the unions, Babcock had noted the Scottish government's plan to remodel Rosyth and Faslane after independence, but said it remained unclear whether workloads would be smaller or greater than now.
Label	**Attack**	**NA**
Reason for Label	In this comment pair, although Comment B concedes one of Comment A's points, it is clear that their overall stance on the main topic under discussion (the obsolescence of aircraft carriers) differs from Comment A, so we count this as an attack	In this pair of comments, the stance of Comment B towards Comment A is not clear (there appear to be both attacking and supporting statements), so we mark it as NA.

Table 14: Comments with Mixed Attack/ Support Relations

Concession-Counterarguments In some cases, an interlocutor may concede the merit of an opponent's argument, but then provide an alternative argument for their side of the debate. Even though they explicitly or implicitly contain a concession, we will still regard them as “attacks" because they attack the primary argument given in the previous comment.

Example	Good point. Against less high tech adversaries, carriers still have merit.
Label	**Attack**
Reason For Label	An argument which concedes on one aspect but nevertheless attacks the main point put forward by the other side

Table 15: Example of the Concession-Counterargument Pattern

Explanation vs Argument Sometimes it may be ambiguous as to whether a comment is “explaining facts" or making an argument. This should be decided by the context of the explanation within the comment thread; in some cases, an explanatory comment is posted merely because the comment thought that the facts that it contained could be of interest to other users, whereas in other cases, an

explanation is posted in order to try and persuade another user to accept his or her perspective.

Example	I am Welsh, I live in Wales. I worry greatly about the version of History taught to us. We do not teach ourselves enough of the ills, violence and grievous nature of Empire, we have plowed on with expansionist economic policy which itself morphed out of empire. [LONG EXPLANATION WITH EXAMPLES] It will not bode well for us in a global world that is changing rapidly, either to continue our pursuit of aggressive foreign policy or to ignore the ills of the past, we must as a culture evolve peacefully or face destruction - that is my view.	**Comment A:** Saying that the floods were hailed as a "once-in-100-years" event, is misleading. 1:100 year flood is a technical way of describing the scale of the flood, and doesn't mean that a similar size event could not occur 2 years later. **Comment B:** The input timescale is almost certainly not 100 years of data, so that the probability is a pure guess; more likely its an extrapolation based on the typical variation in rainfall observed over say 50 years. Given the Earth is 2 billion years old the nature of these assumptions is easy to black-swan-the-fuck-outta-here. Still, its the most useful and workable measure we have.
Label	**Attack**	**NA**
Reason for Label	This is a comment which contains an element of explanation, but since this explanation is ultimately used in service of putting forward and argument, we still count it as argumentative	In the context, Comment B does not appear to be argumentative, but rather an explanation/ clarification of the point made in Comment A, and therefore we mark it as non-argumentative.

Table 16: Explanation vs. Argument

Sarcasm Due to the inherently ambiguous nature of sarcasm and the lack of tonal cues in written text, comments which you suspect may contain sarcasm can be difficult to annotate. Due to this it is advisable (unless it is made very clear by the context) to label possibly-sarcastic comments as non-argumentative.

Example	Great, let the proles squabble amongst themselves which of their number should consider themselves 'rich' (clue anyone who is 'a couple of paychecks away from the street' as they would say in the US, ain't rich), whilst the oligarchs get on with the job	Yep... another 18 years without accelerated warming should prove it for all time. And the ocean surface temps... a waste of time and printers ink
Label	**Attack**	**NA**
Reason for Label	The word "great" is obvious sarcasm, made clear by the rest of the comment	It is not clear whether or not this comment is a serious comment denying global warming, or a sarcastic comment mocking climate change denialists, so we mark it a NA.

Table 17: Sarcastic comments