Can Large Language Models perform Relation-based Argument Mining?

Anonymous ACL submission

Abstract

Argument mining (AM) is the process of automatically extracting arguments, their components and/or relations amongst arguments and components from text. As the number of platforms supporting online debate increases, the need for AM becomes ever more urgent, especially in support of downstream tasks. Relation-based AM (RbAM) is a form of AM focusing on identifying agreement (support) and disagreement (attack) relations amongst arguments. RbAM is a challenging classification task, with existing methods failing to perform satisfactorily. In this paper, we show that general-purpose Large Language Models (LLMs), appropriately primed and prompted, can significantly outperform the best performing (RoBERTa-based) baseline. Specifically, we experiment with two open-source LLMs (Llama-2 and Mistral) with ten datasets.

1 Introduction

Argument mining (AM) is the process of automatically extracting arguments, their components and/or relations amongst arguments and components from natural language text (Lippi and Torroni, 2016; Lawrence and Reed, 2019). The general AM problem can be split into three main tasks: 1) *argument identification*, involving segmenting text into units and determining which are argumentative; 2) *identification of argumentative components*, typically involving classifying claims and/or premises of argumentative text; and 3) *identification of argumentative relations*, aiming at determining how different texts are related within argumentative discourse.

As the number of platforms supporting online debate increases, the need for AM becomes ever more urgent (Lawrence and Reed, 2019). In this paper, we focus on a special form of AM, within the third category, and matching the kind of debate abstractions in platforms such as `kialo.com`, where arguments (textual comments) are connected via *support* or *attack* argumentative relations. Specifically, we will focus on the form of AM framed as the following (binary) *relation-based AM* (RbAM) task (Carstens and Toni, 2015; Cocarascu and Toni, 2017; Cocarascu et al., 2020):[1] *given a pair* (A, B) *of texts A and B, determine whether A attacks or supports B*. For example, take the three arguments, drawn from the Debatepedia/Procon dataset (Cabrio and Villata, 2014), a_1=‘Abortion should be legal’, a_2=‘A baby should not come into the world unwanted’, and a_3=‘Abortion increases the likelihood that women will develop breast cancer’. Here, a_2 can be deemed to support a_1 and a_3 to attack a_1.

RbAM can be used to support several downstream tasks, for example, to gather evidence (Carstens and Toni, 2015), to determine which online arguments are acceptable (Bosc et al., 2016), and to analyse divisive issues about new regulations (Konat et al., 2016). However, it is a challenging task, with different BERT-based models performing reasonably well on some datasets but individual baselines failing to perform well across datasets (Cocarascu et al., 2020; Ruiz-Dolz et al., 2021).

In this paper, we focus on deploying general-purpose LLMs, with appropriate priming and prompting, to address the RbAM task uniformly across several datasets. In doing so we draw inspiration from recent works showing that LLMs perform significantly better than existing baselines on other AM tasks (Chen et al., 2023; Al Zubaer et al., 2023; van der Meer et al., 2022) (see §2). Overall, our contributions are as follows:

- We provide a method for performing RbAM effectively with chat-based LLMs, appropriately, but simply, primed and prompted (see §3).
- We demonstrate empirically, with a wide-ranging evaluation with ten datasets from the literature (see §4), that our LLM-based method for RbAM outperforms the state-of-the-art RoBERTa baseline for RbAM (Ruiz-Dolz et al., 2021) (see §5).

2 Related Work

Relation-Based Argument Mining. The field of RbAM has received significant attention in recent years (Cabrio and Villata, 2018). Hou and Jochim (2017) introduced a Joint Inference model and compared it against baseline methods of logistic regression, attention-based LSTMs, and the EDITS method from Cabrio and Villata (2012), which recognises textual entailment by calculating the distance between arguments.

[1] In (Carstens and Toni, 2015; Cocarascu and Toni, 2017), the task is framed as a ternary classification problem, including a third class *no relation*. Here, we focus on the binary version experimented with in (Cocarascu et al., 2020).

Their method outperformed the baselines with an F_1 score of 65, on the Debatepedia/Procon dataset (Cabrio and Villata, 2014), which we also use (but they do not include the Procon debates). Cocarascu and Toni (2017) used a deep learning architecture with two separate LSTMs on the embeddings of the two arguments in each pair, concatenating the outputs using a softmax layer. Their method achieved an F_1 score of 89 on the Web-Content dataset (Carstens and Toni, 2015) that we also use. Cocarascu et al. (2020) used four deep learning architectures with different types of embeddings and compared them against baselines of Random Forests and SVMs. Their method achieved a best macro F_1 score of 54, which performed similarly to the baselines, on ten datasets, most of which we also use[2]. Another relevant work is by Trautmann et al. (2020), who experimented with several variants of LSTMs, CAM-Bert, and TACAM-BERT on the UKP corpus (Stab et al., 2018) that we also use, achieving a best F_1 score of 80 with TACAM-BERT. Meanwhile, Jo et al. (2021) used Logical Mechanisms and Argumentation Schemes, with, as baselines, TGA Net, Hybrid Net, BERT, BERT+Latent Cross, and BERT+Multi-task Learning. Their best model achieved an F_1 score of 77 with a dataset also collected from the online debate site Kialo as one of our datasets, and an F_1 score of 80 on a similar dataset to Debatepedia/Procon (Cabrio and Villata, 2014) that we use (but without including the Procon debates). Finally, Ruiz-Dolz et al. (2021) evaluated various BERT-based models against LSTMs, achieving an F_1 score of 70 with RoBERTa-large on the US2016 debate corpus and the Moral Maze multi-domain corpus, both from AIFdb (which we do not use – see footnote 2).

None of the mentioned approaches to RbAM use LLMs, nor do they achieve the satisfactory performance across datasets that we aim for.

Argument Mining via Large Language Models. Recently, the exceptional performance of LLMs across a variety of NLP tasks has led to investigations into their performance in a number of AM tasks. Chen et al. (2023) tested the capabilities of LLMs for claim detection, evidence detection, stance detection[3], evidence type classification, and argument generation. They used GPT-3.5-Turbo, Flan-UL2, and Llama 2 13B models for testing, demonstrating that the LLMs perform well in these tasks. Thorburn and Kruger (2022) fine-tuned GPT Neo, a pre-trained LLM, to generate, by prompting, natural language arguments supporting or attacking a topic argument. However, work is still to be done before LLMs can be deemed to reason argumentatively, a finding echoed by Hinton and Wagemans (2023). Further challenges are pointed out by Ruiz-Dolz and Lawrence (2023), who attempted to use LLMs to detect argumentative fallacies but showed that LLMs did not surpass the performance of the RoBERTa-based Transformer model. Meanwhile, Al Zubaer et al. (2023) focused on the classification of argument components in the legal domain with the GPT-3.5 and GPT-4 models, using a bespoke a few-shot prompting strategy, showing that the LLMs did not surpass the domain-specific BERT-based baseline. More promising results were found in a study of LLMs' potential for generating counter-narratives to counteract online hate speech when supplemented by argumentative strategies and analysis (Furman et al., 2023). Here, the argumentative information, provided by either fine-tuning or priming, was shown to improve the quality of the generated counter-narratives in both English and Spanish. LLMs' potential for AM was also seen by van der Meer et al. (2022), who used LLMs for argument quality prediction, amounting to classifying the validity and novelty of a given argument, comprising a premise and a conclusion. They achieved best performance using a few-shot learning priming strategy with LLMs for the validity task and a Transformer-based model fine-tuned for the novelty task.

Importantly, to the best of our knowledge, no study to date considered the use of LLMs for RbAM.

3 LLMs for RbAM

Our method is overviewed in Figure 1. It consists of few-shot priming, which has shown to perform well with LLMs without the need for fine-tuning (Brown et al., 2020), followed by prompting. The primer uses four labelled examples of attack and support relations between arguments, before we provide an example in the prompt for the LLM to classify as attack or support. The four examples in the primer are fixed text comprising a parent argument (Arg1), a child argument (Arg2) and the classification of the relation from the child to the parent argument, as shown in the top, pink part of the box in Figure 1. Then, the prompt amounts to a pair of arguments presented as the four in the primer, but without indicating the relation, as shown in the bottom, turquoise part of the box in Figure 1. In the experiments, the parent and child arguments in the prompt are inputs (from the RbAM datasets described in §4). Examples of some of these prompts are given in Appendix A.

4 Experimental Set-up

We describe the datasets used, the baseline we compare against and the LLMs we experiment with.[4]

Datasets We used ten existing datasets, as follows (see Appendix B for additional information, including statistics). Note that the datasets labelled * directly fit the RbAM task definition (classification of pairs of texts). The dataset labelled † is an extension of a dataset

[2]We do not use AIFdb (`https://corpora.aifdb.org/`) as it is not obvious how to map it univocally onto RbAM.

[3]This deals with classifying the stance of arguments towards topics, whereas RbAM deals with classifying the relation between (two) arguments.

[4]All our experiments are executed with two RTX 4090 24GB on an Intel(R) Xeon(R) w5-2455X. In total, it took 112.3 hours to run all the LLM experiments.

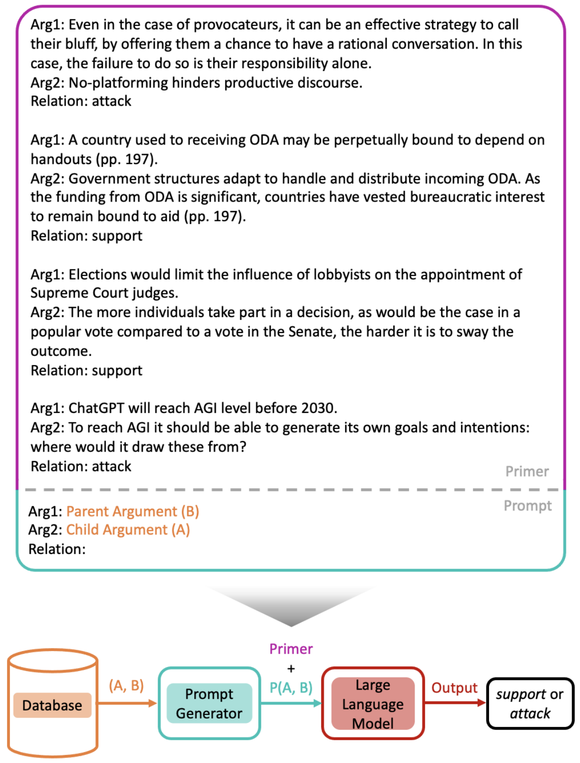

Figure 1: Experimental pipeline with the (few-shot learning) primer and the prompt template P(A,B).

already fitting the RbAM task definition to include additional relations between sentences and topics. For all these RbAM datasets, we have ignored any relations other than attack and support, given our focus on binary RbAM. The other datasets are originally given for different tasks, e.g. to determine relations between sentences and topics or between premises and claims: we adapt them to the RbAM task as discussed in the following.

Persuasive essays (Essay) (Stab and Gurevych, 2017) is a corpus of annotated 402 persuasive essays.

Microtexts (Mic)* (Peldszus and Stede, 2015) is a corpus of 112 short texts on controversial issues, with 576 arguments. They were originally written in German and then translated to English.

Nixon-Kennedy debate (NK)* (Menini et al., 2018) is a corpus from the 1960 Nixon-Kennedy presidential campaign covering five topics.

Debatepedia-Procon (DP)* (Cabrio and Villata, 2014) is a corpus extracted from two online debate platforms: Debatepedia and Procon.

IBM-Debater (IBM) (Bar-Haim et al., 2017) is a dataset containing 55 controversial topics collected from the debate motions database at the International Debate Education Association (IDEA) website.

ComArg† (Boltužić and Šnajder, 2014) is a corpus of user comments collected from Procon and IDEA where each argument has a stance for or against one of two topics. For our experiments we adapted the dataset so that the parent argument is the topic. Also, we set explicit and vague/implicit attacks to be attacks and vague/implicit and explicit supports to be supports.

*CDCP** (Park and Cardie, 2018) is a corpus annotated with only support relations containing 731 user comments on Consumer Debt Collection Practices from the eRulemaking platform.

UKP (Stab et al., 2018) is a corpus with arguments obtained from Web documents (including news reports, editorials, blogs, debate forums, and encyclopedias) over eight controversial topics. We adapted the parent argument to be '*topic* is good' (e.g. 'abortion is good', where abortion is one of the topics).

Web-Content (Web)* (Carstens and Toni, 2015)[5] contains arguments adapted from the Argument Corpus (Walker et al., 2012), plus arguments from news articles, movies, ethics and politics.

*Kialo** was collected from the online debate platform Kialo. Debates (in English) were scraped from Kialo (in 2022) on topics related to Politics, Law, and Sports.

Baseline We opted to fine-tune **RoBERTa**, given its performances in (Ruiz-Dolz et al., 2021). We fine-tuned it with 75% of each dataset separately for 50 epochs (25% of the datasets were kept for validation), a batch size of 8, and a learning rate of 1e-5. For each dataset, we selected the best model (over the 50 epochs), i.e. that which achieved the highest F_1 score on the validation set. We then used these candidate models (one for each dataset) to perform inference for the other datasets and selected the best (which turned out to be the one trained on Kialo) as the baseline (for performances of all these models see Appendix C).

Large Language Models. We chose two families of LLMs, both open-source (details are in Appendix D). Since LLMs have a huge number of parameters and require a large amount of GPU space, there have been attempts to reduce the space they take by compressing them to smaller sizes. For example, GPTQ (Frantar et al., 2022) uses one-shot weight quantisation based on approximate second-order information to reduce the bit size of each weight in the LLM. So, for all three LLMs considered, we also experimented with 4bit quantisation (so each weight is stored in 4bits on the GPU) as it had the best trade-off between accuracy and space.

The **Llama 2** models (Touvron et al., 2023) have been pre-trained with 2 trillion tokens and are generally good at causal language modelling. In our experiments, we used the Llama 2 13B model (and its GPTQ quantised version) which has 13 billion parameters and the Llama 2 70B (GPTQ quantised as the base model needs nearly 140GB of GPU space) which has 70 billion parameters.

The **Mistral 7B** model (Jiang et al., 2023) is a 7 billion parameter pre-trained and fine-tuned LLM. The model is claimed to perform better than any other open source 13 billion parameter LLM (including Llama 2 13B) (Jiang et al., 2023). The **Mixtral 8x7B** model (Jiang et al., 2024) builds on the Mistral 7B model by

[5]To access the dataset, see: https://www.doc.ic.ac.uk/~oc511/ACMToIT2017_dataset.xlsx

	RoBERTa	Llama13B	Llama13B-4bit	Llama70B-4bit	Mistral7B	Mixtral-8x7B-4bit
Essays	85 / 38 / 80	87 / 31 / 82	91 / 36 / 86	94 / 52 / **90**	89 / 42 / 85	94 / 43 / 89
Nixon-Kennedy	56 / 67 / 62	67 / 12 / 39	66 / 5 / 34	64 / 71 / **68**	54 / 68 / 61	66 / 50 / 58
CDCP	75 / - / 75	87 / - / 87	94 / - / **94**	92 / - / 92	75 / - / 75	93 / - / 93
UKP	68 / 81 / 75	70 / 82 / 77	75 / 84 / 80	84 / 89 / **87**	78 / 83 / 81	81 / 84 / 83
Debatepedia/Procon	90 / 89 / 90	83 / 71 / 77	84 / 72 / 79	96 / 95 / **96**	90 / 89 / 90	94 / 93 / 94
IBM-Debater	85 / 82 / 83	81 / 66 / 75	88 / 82 / 85	94 / 92 / 93	89 / 89 / 89	95 / 93 / **94**
ComArg	71 / 74 / 72	68 / 62 / 65	70 / 58 / 65	77 / 56 / 68	56 / 71 / 63	79 / 73 / **76**
Microtexts	73 / 53 / 67	76 / 45 / 67	84 / 41 / 72	81 / 52 / **73**	71 / 54 / 67	80 / 45 / 70
Web-Content	67 / 67 / 67	66 / 63 / 64	68 / 53 / 60	72 / 72 / **72**	57 / 72 / 64	70 / 66 / 68
Kialo	- / - / -	74 / 56 / 65	75 / 54 / 65	87 / 84 / **86**	83 / 83 / 83	85 / 82 / 84
Average	74 / 61 / 75	76 / 49 / 70	79 / 48 / 72	84 / 66 / **82**	74 / 65 / 76	84 / 63 / 81
Macro F_1	68	62	64	**75**	70	73
Inference Time (s)	0.005	0.11	0.34	1.73	0.06	0.28

Table 1: F_1 scores (as a percentage) for support / attack / both relations in various datasets (rows) for the models used (columns). RoBERTa here is the baseline (see §4) and boldface font indicates the best performing model (for both relations) for each dataset. The last row gives the time it takes for a single inference for each model, in seconds.

using 8 of them: for every token, the model selects two of the Mistral 7B models to produce an output and combines them (Jiang et al., 2024). Its performance is claimed to be equal to the Llama 2 70B model (Jiang et al., 2024). In our experiments, we used the Mistral 7B model and the Mixtral 8x7B model (GPTQ quantised as the base model needs nearly 95GB of GPU space).

5 Results

Table 1 shows the results.[6] We can see that Llama 70B-4bit achieved the highest macro F_1 score of 75, outperforming all of the baselines. Also, in seven of the datasets (Essay, NK, UKP, DP, Mic, Web, and Kialo), it achieved the highest F_1 score of all LLMs (as well as better than all baselines in all of these datasets except two, see Appendix C). However, the inference time of 1.73 seconds per argument pair for this model was rather high (we believe this is not just because it is the biggest model, but also because it is GPTQ quantised).

Mixtral 8x7B-4bit performed almost as well as Llama 70B-4bit, with a macro F_1 score of 73, with average F_1 score for the support labels as for Llama 70B-4bit but the average F_1 score for the attack labels 3 points lower. However, it achieved the highest F_1 scores in two datasets (IBM and ComArg). Its inference time was (a much lower) 0.28 seconds per argument pair (we believe it may be faster still if we did not use quantisation).

Mistral 7B performed well given that it is smaller than the other LLMs used, achieving a macro F_1 score of 70 which was better than any of the baselines (see Appendix C). However, it did not outperform other LLMs in any dataset. Mistral 7B was also the fastest, with an inference time of 0.06 seconds per argument pair.

[6]In the vast majority of cases, the LLMs responded with either *attack* or *support*, as expected. However, for 43 instances the LLMs generated other labels (see Appendix E), a very small number in comparison with the total number of pairs assessed (159604): we ignored them in the results.

Llama 13B and Llama 13B-4bit achieved similar macro F_1 scores, 62 and 64, respectively. However, their performance on each dataset was varied. Llama 13B-4bit performed best on CDCP, which was expected as CDCP only contains support labels and Llama 13B-4bit tends to output support more often. Note that, with GPTQ quantisation, the performance improves. They both performed worse than the best baselines (see Appendix C). We note that Llama13B-4bit was unexpectedly slower than Llama13B.

6 Conclusion and Future Work

We have introduced a method for the RbAM task using general purpose LLMs, appropriately primed and prompted. We showed, with experiments on ten datasets and five open-source LLMs (more than half of which quantised), that **Llama 70B-4bit** and **Mixtral 8x7B-4bit** surpassed the RoBERTa baseline, with the former outperforming the latter but also bringing the downsides of slower inference time and greater GPU requirements.

For future work there are many potential avenues, including the following three: 1) We could mask the entities in sentences to outline their argumentative structure, which is shown to improve performance for the argument retrieval task (Ein-Dor et al., 2020). 2) We plan to work on improving the prediction on the attack relations as LLMs and also baselines performed worse on them. 3) We plan to extend this work for the more challenging (ternary) RbAM task, i.e. determining whether there is a support, an attack or no relation between two arguments.

7 Limitations

There are some limitations of our work. First, the task that we consider is the (binary) RbAM task (identifying support/attack) whereas, in most real-world applications, it would be a (ternary) RbAM task (identifying support/attack/no relation) as we discussed in §6. Further, the datasets we used are in English: we are not sure if LLMs will perform as well on RbAM in other languages. GPU limitations affect our selection of small/quantised models, and we were not able to fine-tune any of the LLMs as it was computationally infeasible.

8 Ethics Statement

There are potential risks of LLMs such as social bias and generation of misinformation. In this work, we only use LLMs to generate a single token which is support/attack, so there are no risks of generating biased or false information.

References

Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. 2023. Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, 6.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Tweeties squabbling: Positive and negative results in applying argument mining on social media. 287:21–32.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2014. Node: A benchmark of natural language arguments. In *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 449–450. IOS Press.

Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5427–5433, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. In *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 45–52. IOS Press.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7683–7691. AAAI Press.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: accurate post-training quantization for generative pre-trained transformers.

Damián Ariel Furman, Pablo Torres, José A. Rodríguez, Diego Letzen, Maria Vanina Martinez, and Laura Alonso Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, Singapore, December 6-10, 2023, pages 2942–2956. Association for Computational Linguistics.

Martin Hinton and Jean H. M. Wagemans. 2023. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument Comput.*, 14(1):59–74.

Yufang Hou and Charles Jochim. 2017. Argument Relation Classification Using a Joint Inference Model. In *Proceedings of the 4th Workshop on Argument Mining*, pages 60–66, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. 2016. A corpus of argument networks: Using graph properties to analyse divisive issues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3899–3906, Portorož, Slovenia. European Language Resources Association (ELRA).

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10:1–10:25.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. pages 4889–4896.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barberá, and Ana García-Fornes. 2021. Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation. *IEEE Intelligent Systems*, 36(6):62–70. Conference Name: IEEE Intelligent Systems.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models. In *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Luke Thorburn and Ariel Kruger. 2022. Optimizing language models for argumentative reasoning. In *Proceedings of the 1st Workshop on Argumentation & Machine Learning co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 13th, 2022*, volume 3208 of *CEUR Workshop Proceedings*, pages 27–44. CEUR-WS.org.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Dietrich Trautmann, Michael Fromm, Volker Tresp, Thomas Seidl, and Hinrich Schütze. 2020. Relational and Fine-Grained Argument Mining: The LMU Munich project ReMLAV within the DFG Priority Program RATIO "Robust Argumentation Machines". *Datenbank-Spektrum*, 20(2):99–105.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Appendix

A Example Prompts

In this section we give example prompts generated from each dataset (except the Kialo and UKP datasets as these datasets do not allow us to share them), as seen from Figures 2,3,4,5,6,7,8,9.

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: using machines is advantageous
Arg2: the usage of machines is harmful for health of humans
Relation:

Figure 2: An example prompt drawn from the Essays dataset used in the RbAM experiments.

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: The death penalty should be abandoned everywhere.
Arg2: Moreover it turns out time and again that innocent people are also convicted and executed.
Relation:

Figure 3: An example prompt drawn from the Microtexts dataset used in the RbAM experiments.

B Datasets

Number of support/attack relations for all these datasets are given in Table 2. This information is important when

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: 11 . Kennedy 's statement : The Republicans have consistently opposed minimum wage legislation . Fact : In 1938 , when the first bill was passed , the Republicans voted against it 48 to 31 in the House and 13 to 2 in the Senate .
Arg2: Mr. Nixon voted against it , every single time , and I voted for it . Minimum wage - I see some signs waved around by great supportcrs of Mr. Nixon . I want to ask them three questions .
Relation:

Figure 4: An example prompt drawn from the Nixon-Kennedy dataset used in the RbAM experiments.

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: Abortion should be legal
Arg2: A baby should not come into the world unwanted
Relation:

Figure 5: An example prompt drawn from the Debatepedia/Procon dataset used in the RbAM experiments.

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: This house believes all nations have a right to nuclear weapons
Arg2: public opinion is overwhelmingly opposed to nuclearization
Relation:

Figure 6: An example prompt drawn from the IBM-Debater dataset used in the RbAM experiments.

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: Gay marriage should be legal.
Arg2: It is discriminatory to refuse gay couples the right to marry
Relation:

Figure 7: An example prompt drawn from the ComArg dataset used in the RbAM experiments.

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: However, I don't think the law, as written, is easy to understand.
Arg2: I think the law should be clarified,
Relation:

Figure 8: An example prompt drawn from the CDCP dataset used in the RbAM experiments.

Arg1: Even in the case of provocateurs, it can be an effective strategy to call their bluff, by offering them a chance to have a rational conversation. In this case, the failure to do so is their responsibility alone.
Arg2: No-platforming hinders productive discourse.
Relation: attack

Arg1: A country used to receiving ODA may be perpetually bound to depend on handouts (pp. 197).
Arg2: Government structures adapt to handle and distribute incoming ODA. As the funding from ODA is significant, countries have vested bureaucratic interest to remain bound to aid (pp. 197).
Relation: support

Arg1: Elections would limit the influence of lobbyists on the appointment of Supreme Court judges.
Arg2: The more individuals take part in a decision, as would be the case in a popular vote compared to a vote in the Senate, the harder it is to sway the outcome.
Relation: support

Arg1: ChatGPT will reach AGI level before 2030.
Arg2: To reach AGI it should be able to generate its own goals and intentions: where would it draw these from?
Relation: attack

Primer

Prompt

Arg1: Transparency is necessary for security
Arg2: Transparency can result in normalisation
Relation:

Figure 9: An example prompt drawn from the Web-Content dataset used in the RbAM experiments.

the F_1 scores are calculated. Also, when RoBERTa is fine-tuned on these datasets it is important point how balanced the datasets are.

Datasets	#Support	#Attack	Total#
Essays	4841	497	5338
Microtexts	322	121	443
Nixon-Kennedy	356	378	734
Debatepedia/Procon	319	261	580
IBM-Debater	1325	1069	2394
ComArg	640	484	1124
CDCP	1284	0	1284
UKP	4944	6195	11139
Web-content	1348	1316	2664
Kialo	68549	65355	133904

Table 2: Number of support/attack relations in each dataset.

Number of average words and characters for each dataset are given in Table 3. This kind of statistics help with understanding why all the models underperformed on a specific dataset. For example, in the Nixon-Kennedy dataset the average argument is very long with 103.57 words per argument which contains a lot more information for any model to process and it can be seen that the accuracy is lacking.

Datasets	Average # of words	Average # of characters
Essays	14.7	87.09
Microtexts	13.58	81.3
Nixon-Kennedy	103.57	539.21
Debatepedia/Procon	34.81	215.22
IBM-Debater	10.78	68.84
ComArg	56.81	318.55
CDCP	15.4	88.11
UKP	15.33	83.64
Web-content	19.87	112.94
Kialo	21.84	135.69

Table 3: Statistical features of each dataset.

C RoBERTa Baselines

Table 4 shows the results for the baselines in the RbAM task, i.e. RoBERTa fine-tuned on each dataset and then evaluated on the remaining datasets.

RoBERTa fine-tuned with the Kialo dataset achieved the highest macro F_1 score of 68 and an F_1 score better than other baselines in four datasets (NK, UKP, and Web). However, note that, since the dataset is large it took a long time to fine-tune, specifically 53.73 hours.

RoBERTa fine-tuned with the DP and the IBM datasets both achieved a macro F_1 score of 66, which came close to the RoBERTa fine-tuned with the Kialo dataset. RoBERTa fine-tuned with the DP dataset achieved a better F_1 score than other baselines in three datasets (ComArg, Mic, and Kialo). These datasets are smaller than Kialo and so fine-tuning took 0.23 hours for the DP dataset and 0.96 hours for the IBM dataset.

We thus selected RoBERTa fine-tuned with the Kialo dataset as the best baseline, as it performed better than other baselines. We note here also that for all of the baseline models, a single inference took 0.005 seconds for each test sample.

D LLMs

The amount of GPU space needed for Llama 13B is 27GB, Llama 13B-4bit is 7.4GB, Llama 70B-4bit is 37GB, Mistral 7B is 15GB, and Mixtral 8x7B-4bit is 25GB. For every model, we use the default parameter selection for temperature=0.7, top_p=1, do_sample=False. However, max_new_tokens=1 as inference time is faster and we only need a single token generated for support/attack. Also, the models that are not quantised are loaded with 16-bit precision for faster inference.

E Extra labels

Across the datasets, there were 43 instances where the LLMs generated additional labels than attack/support. The additional labels the LLMs generate are different for all the models, as shown in Table 5.

	Essay	NK	CDCP	UKP	DP	IBM	ComArg	Mic	Web	Kialo
Essay	- / - / -	95 / 5 / 86	95 / 0 / 86	71 / 25 / 67	90 / 42 / 85	89 / 41 / 84	94 / 45 / **90**	79 / 14 / 73	56 / 16 / 52	85 / 38 / 80
NK	65 / 0 / 32	- / - / -	65 / 0 / 32	54 / 46 / 50	65 / 31 / 47	60 / 55 / 58	65 / 4 / 34	64 / 1 / 32	46 / 48 / 47	56 / 67 / **62**
CDCP	1 / - / **1**	98 / - / 98	- / - / -	42 / - / 42	90 / - / 90	77 / - / 77	98 / - / 98	95 / - / 95	34 / - / 34	75 / - / 75
UKP	67 / 42 / 53	61 / 28 / 43	61 / 0 / 27	- / - / -	68 / 75 / 72	73 / 75 / 74	74 / 67 / 70	51 / 47 / 49	58 / 38 / 47	68 / 81 / **75**
DP	75 / 34 / 57	72 / 23 / 50	71 / 0 / 39	62 / 67 / 64	- / - / -	84 / 82 / 83	85 / 78 / 82	71 / 0 / 39	61 / 43 / 53	90 / 89 / **90**
IBM	76 / 37 / 59	72 / 26 / 51	71 / 0 / 39	58 / 69 / 63	82 / 78 / 80	- / - / -	87 / 83 / 85	60 / 33 / 48	68 / 17 / 45	85 / 82 / 83
Com-Arg	76 / 36 / 59	72 / 2 / 42	73 / 0 / 41	59 / 62 / 60	82 / 73 / **78**	73 / 71 / 72	- / - / -	72 / 5 / 43	72 / 3 / 42	71 / 74 / 72
Mic	85 / 28 / 70	83 / 3 / 61	84 / 0 / 61	52 / 44 / 50	83 / 51 / **74**	77 / 52 / 71	83 / 33 / 69	- / - / -	60 / 34 / 53	73 / 53 / 67
Web	68 / 13 / 41	67 / 15 / 41	67 / 0 / 34	51 / 67 / 59	65 / 59 / 62	65 / 60 / 63	69 / 32 / 51	61 / 40 / 51	- / - / -	67 / 67 / 67
Kialo	70 / 18 / 45	68 / 14 / 42	68 / 0 / 35	46 / 63 / 54	79 / 71 / **75**	74 / 73 / 73	74 / 52 / 63	67 / 3 / 36	61 / 36 / 49	- / - / -
Avg.	76 / 23 / 57	76 / 13 / 57	73 / 0 / 44	55 / 49 / 57	78 / 53 / 74	75 / 57 / 73	81 / 44 / 71	69 / 16 / 52	57 / 26 / 47	74 / 61 / **75**
Mac. Avg.	0.50	0.45	0.36	0.52	0.66	0.66	0.62	0.42	0.42	**0.68**
Train Time (in hours)	2.14	0.29	0.52	4.47	0.23	0.96	0.45	0.18	1.07	53.73

Table 4: F_1 scores for various datasets (rows) by the RoBERTa baselines, fine-tuned on the datasets (columns), where F_1-S stands for the F_1 score of the *support* relation, F_1-A stands for the F_1 score of the *attack* relation and boldface font indicates the best performing baseline for each dataset. The training time it takes for each RoBERTa model, fine-tuned on the datasets is given in hours in the last row.

	Llama 13B	Llama 13B-4bit	Llama 70B	Mistral	Mixtral
Kialo	compare (25) conflict (1)	compare (1)		analogy (1)	irrelevant (2) contradiction (2) compare(2) contrast(1)
Essays			paraphrase (1)		contradiction (1)
UKP					contradiction (1)
Web	reply (1)				
ComArg					paraphrase (1)
CDCP					paraphrase (1)
NK	rebuttal (2)				

Table 5: These are the additional labels the LLMs generated (columns) on the datasets (rows). The number in the parentheses represents the number of times the label has been generated.