# Can Large Language Models perform Relation-based Argument Mining?

**Deniz Gorur and Antonio Rago and Francesca Toni**
Department of Computing, Imperial College London, UK
{d.gorur22,a.rago,ft}@imperial.ac.uk

## Abstract

Relation-based Argument Mining (RbAM) is the process of automatically determining agreement (support) and disagreement (attack) relations amongst textual arguments (in the binary prediction setting), or neither relation (in the ternary prediction setting). As the number of platforms supporting online debate increases, the need for RbAM becomes ever more urgent, especially in support of downstream tasks. RbAM is a challenging classification task, with existing state-of-the-art methods, based on Language Models (LMs), failing to perform satisfactorily across different datasets. In this paper, we show that general-purpose Large LMs (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) and with GPT-3.5-turbo on several datasets for (binary and ternary) RbAM, as well as with GPT-4o-mini on samples (to limit costs) from the 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 specific form of AM, within the third category, and matching the kind of debate abstractions in platforms such as Kialo[1] and ArguCast[2] (Gorur et al., 2023), 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 (ternary) *relation-based AM* (RbAM) task (Carstens and Toni, 2015; Cocarascu and Toni, 2017; Cocarascu et al., 2020):[3]

> given a pair $(A, B)$ of texts $A$ and $B$, determine whether $A$ attacks, supports, or has neither relation to $B$.

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'. In this example , $a_2$ can be deemed to support $a_1$, $a_3$ to attack $a_1$, and $a_2$ as being in neither relation with $a_3$.

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

[1] www.kialo.com

[2] www.argucast.herokuapp.com/

[3] We experiment with both the binary version (without the *neither* label, also experimented with in (Cocarascu et al., 2020)) and ternary version (as in the definition given).



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534
January 19–24, 2025. 

perform significantly better than existing baselines on other AM tasks (Chen et al., 2024; Zubaer et al., 2023; van der Meer et al., 2022) (see §2).

Overall, our contributions are as follows: We provide a novel 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 on eleven datasets from the literature (see §4), that our LLM-based method for RbAM outperforms the state-of-the-art RoBERTa baseline for binary RbAM (Ruiz-Dolz et al., 2021) (see §5.1). We also demonstrate empirically, on four of the datasets, that LLMs for RbAM outperforms state-of-the-art RoBERTa baseline for ternary RbAM (see §5.2). Further, we show that our approach is robust to different primers in ablation studies (see §6) and that it performs well on the recently proposed ARIES benchmark (Gemechu et al., 2024) (see §7).

The code for all experiments is available at github.com/dnzggg/Can-LLMs-Perform-RbAM.

## 2 Related Work

### 2.1 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. 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 note that we include the Procon *debates* that they exclude). 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[4]. Another relevant work is by Trautmann et al. (2020), who experimented with several variants of LSTMs, CAM-Bert, and TACAM-BERT, achieving the best $F_1$ score of 80 on the UKP corpus (Stab et al., 2018) that we also use. Meanwhile, Jo et al. (2021) used Logical Mechanisms and Argumentation Schemes, with baselines such as 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, again, we include the Procon debates that they excluded). 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[4]). Finally, the recently introduced ARIES benchmark (Gemechu et al., 2024) consists of eight diverse datasets and three LM-based baselines for RbAM (DialoGPT, T5, and RoBERTa, which we use as our baseline), we consider these datasets for our experiments (see §7).

Overall, while advancements in RbAM have clearly been made, our aim is to explore whether LLMs' linguistic abilities can set the bar higher in this task.

### 2.2 Argument Mining via LLMs

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. (2024) tested the capabilities of LLMs on: claim detection, evidence detection, stance detection[5], evidence type classification, and argument generation. They used GPT-3.5-turbo, Flan-UL2, and Llama2-13B models, 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. 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 prim-

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

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

ing, 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 for the validity task and a Transformer-based model fine-tuned for the novelty task.

Despite these successes, 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, 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 and showing that the LLMs did not surpass the domain-specific BERT-based baseline. Recently, Saadat-Yazdi and Kökciyan (2024) explored the use of LLMs for argument canonicalisation, which involves rephrasing arguments into standard forms. They found that LLMs underperform compared to humans.

Importantly, to the best of our knowledge, no study to date considered the use of LLMs for RbAM, which we study in this paper.

## 3 LLMs for RbAM

Our method for utilising LLMs to tackle RbAM is outlined in Figure 1. We adopt a modular approach that includes few-shot priming, which has been shown to perform well with LLMs without the need for fine-tuning (Brown et al., 2020), followed by the task definition and prompting.

The primer includes labelled examples of attack, support, and, for ternary RbAM, neither relations between arguments, followed by an unlabelled example in the prompt for the LLM to classify as attack, support, or, for ternary RbAM, neither. The primer can be adjusted to include $n \in \mathbb{Z}^+$ attack, $p \in \mathbb{Z}^+$ support, and $q \in \mathbb{Z}^+$ neither relation examples. We will refer to different primers by the number of attack/support/neither examples, using the format nApSqN (e.g. in our main ternary RbAM experiments, we use one attack, support, neither examples, denoted as 1A1S1N). For the experiments, the primers were fixed and were randomly drawn from primer seeds from the datasets. The labelled examples in the primer comprise of a parent argument (Arg1), a child argument (Arg2), and the classification of the relation from the child to the parent argument as either attack, support, or, for ternary RbAM, neither, as shown in the top, pink part of the rounded rectangle in Figure 1. The task definition explains what the LLMs should output and the definitions for each output, as shown in the middle, yellow part of the rounded rectangle in Figure 1. Finally, the prompt consists of a pair of arguments presented as the examples in the primer, but without indicating the relation, as shown in the bottom, turquoise part of the rounded rectangle in Figure 1. Here, blue text was added only if the task was ternary RbAM.[6] We conducted preliminary experiments on a subset of the datasets to decide which primer to use, as well as whether to include the task definition and instruction template (for the open-source models) (see Appendix A for the preliminary experiments).

For the main binary RbAM experiments we used a 2A2S-shot primer and for the ternary RbAM experiments we used a 1A1S1N-shot primer. Additionally, the open-source models support the use of an instruction template ([INST]. . . [/INST]) within the prompt. We conducted preliminary experiments to determine whether incorporating the instruction template into LLMs was beneficial.

## 4 Experimental Set-up

In this section, we describe the datasets used (§4.1); the baseline we compare against (§4.2); and the LLMs we experiment with (§4.3).[7]

### 4.1 Datasets

We used eleven existing datasets, as follows (see Appendix D for additional information, including statistics). Note that the datasets labelled * are those where we fully adapted the dataset to fit the RbAM task. The dataset labelled † is an extension of a dataset already fitting the RbAM task definition to include additional relations between sentences and topics. These datasets were originally given for different tasks, such as determining relations between sentences and topics or between premises and claims. We adapted them to fit the RbAM task, as will be discussed below.

[6] Some example prompts are given in Appendix C.

[7] All our experiments are executed with two RTX 4090 24GB on an Intel(R) Xeon(R) w5-2455X.

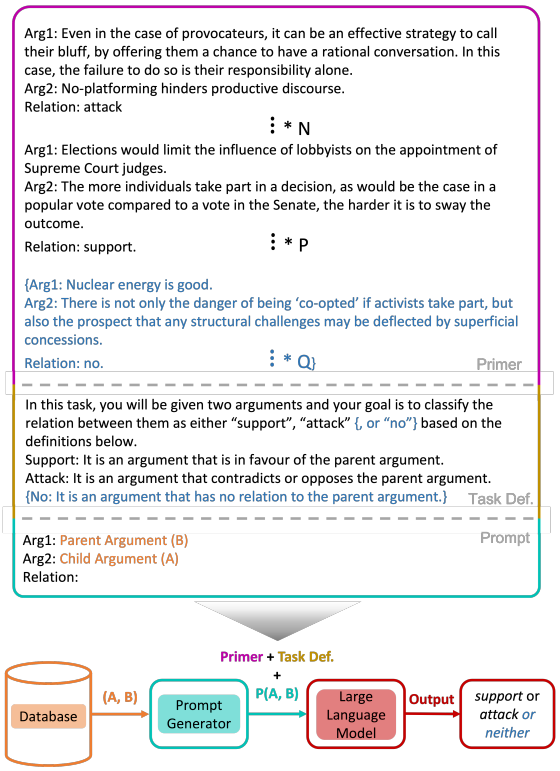


Figure 1: Experimental (modular) pipeline for RbAM with the few-shot learning primer, task definition, and the prompt template P(A,B). The blue text is added for ternary RbAM. The first and second examples are from the Kialo dataset, and the third example is from the UKP dataset. (The examples in the primers were not used in our experiments.)

Some of these RbAM datasets only contain support/attack relations and are therefore suitable only for the binary RbAM task. Other datasets, which contain support/attack/neither relations, can be used for both RbAM tasks (ignoring any relations other than attack and support for binary). Datasets that are suitable for both binary and ternary RbAM tasks are labelled ‡.

Table 1 shows the number of support, attack, neither relation instances in each dataset. This information is important when the $F_1$ scores are calculated, especially for the baseline, as when RoBERTa is fine-tuned on these datasets one can easily see how balanced the datasets are.

*Persuasive essays* (Essay)* (Stab and Gurevych, 2017) is a corpus of 402 persuasive essays annotated with argumentation structures. The initial dataset is composed of major claims that embody the author's viewpoint on the topic, which is then either supported or attacked by claims, which in turn can be supported or attacked by premises. For the purpose of the RbAM task, we have utilised the text from the claims, major claims, and premises.

| Datasets | #Support | #Attack | #Neither | Total# |
|---|---|---|---|---|
| Essays* | 4841 | 497 | - | 5338 |
| Microtexts | 322 | 121 | - | 443 |
| Nixon-Kennedy‡ | 356 | 378 | 1173 | 1907 |
| Debatepedia/Procon | 319 | 261 | - | 580 |
| IBM-Debater* | 1325 | 1069 | - | 2394 |
| ComArg† | 640 | 484 | - | 1124 |
| CDCP | 1284 | 0 | - | 1284 |
| UKP*‡ | 4944 | 6195 | 14353 | 25492 |
| Web-content‡ | 1348 | 1316 | 1394 | 4058 |
| M-Arg‡ | 384 | 120 | 3600 | 4104 |
| Kialo | 68549 | 65355 | - | 133904 |

Table 1: Number of support/attack/neither relations in each dataset. A dash (-) in the #Neither column indicates that the dataset does not contain the neither label.

*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 have been professionally translated to English.

*Nixon-Kennedy debate‡ (NK)* (Menini et al., 2018) is a corpus from the 1960 Nixon-Kennedy presidential campaign covering five topics: Cuba, disarmament, healthcare, minimum wage, and unemployment.

*Debatepedia-Procon (DP)* (Cabrio and Villata, 2014) is a corpus extracted from two online debate platforms: Debatepedia[8] and Procon[9], where users of both systems discuss a set of topics highlighting whether their arguments are in favour of (support) or against (attack) the topic or other users' arguments.

*IBM-Debater* (IBM)* (Bar-Haim et al., 2017) is a dataset containing debates from 55 controversial topics. These debates have been collected from the debate motions database at the International Debate Education Association (IDEA) website[10]. The dataset includes topic texts and claims that support or attack them. For the RbAM task, we used both the topic texts and their claims.

*ComArg†* (Boltuzic and Snajder, 2014) is a corpus of user comments collected from Procon[9] and IDEA[10] platforms. It contains comments on two topics, "Under God in Pledge" and "Gay Marriage", where each argument has a stance for or against one of two topics. For our experiments, we extended the dataset so that the parent argument is the topic (predefined arguments as named in Boltuzic and Snajder (2014)), which can be either "The words 'under god' should be in the U.S. pledge of Allegiance" or "Gay marriage should be legal". Also,

[8] https://idebate.net/resources/debatabase
[9] https://www.procon.org/
[10] https://idebate.net/

we set both explicit and vague/implicit attacks to be attacks and both 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 of arguments obtained from Web documents (including news reports, editorials, blogs, debate forums, and encyclopedias) over eight controversial topics: abortion, cloning, death penalty, gun control, marijuana legalisation, minimum wage, nuclear energy, and school uniforms. We adapted the parent argument for each topic to follow the format '*topic* is good' (e.g. for the topic of abortion, the parent argument would be 'abortion is good').

*Web-Content‡ (Web)* (Carstens and Toni, 2015)[11] contains arguments, adapted from the Argument Corpus (Walker et al., 2012) extracted from an online debate platform. It also includes arguments from news articles, movies, debates of ethics, and politics.

*M-Arg‡* (Mestre et al., 2021) is a multimodal dataset for argument mining, sourced from the US 2020 presidential debates and annotated through crowd-sourcing. While the dataset includes both audio and text, our study in this paper is concerned with text-based LLMs. The dataset covers a wide range of eighteen different topics from the debates, some of the most common topics include COVID, Racism, Climate change, and Economy.

*Kialo* is a dataset collected from the online debate platform Kialo. Debates were scraped from Kialo (in 2022) covering topics related to Politics, Law, and Sports, and then filtered to include the English-only debates.

### 4.2 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), using 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 on the other datasets and selected the best (which turned out to be the one trained on the DP dataset) as the baseline (for performances of all these models, see §5.1 and §5.2).

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

### 4.3 Large Language Models

We chose three families of LLMs, where two were open-source (in the sense that their architectures and parameters are publicly available). Given the large number of parameters and GPU space required by LLMs, there have been attempts to reduce their size by compressing them. Bitsandbytes quantisation (Dettmers et al., 2023) is one such technique that reduces the bit size of each weight in the LLM. So, for all three open-source LLMs considered, we experimented with 4bit quantisation (so each weight is stored in 4 bits on the GPU).

For all models, we constrained the output to the labels we consider for the task. Hyperparameters for every model are set to the default selection of temperature=0.7, top_p=1, do_sample=False, and max_new_tokens=10, except for the closed-source model where max_new_tokens=1.

**Llama 2 Model** 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 decided to use the **Llama2-70B** (4bit quantised, as the base model needs nearly 140GB of GPU space), which has 70 billion parameters and is the best performing Llama 2 model.

**Mistral.AI Models** The **Mistral-7B** model (Jiang et al., 2023) is a 7 billion parameter pre-trained and fine-tuned LLM. It is claimed that this model performs better than any other open-source LLM with 13 billion parameters, including the Llama2-13B model (Jiang et al., 2023).

The **Mixtral-8x7B** model (Jiang et al., 2024) builds on the Mistral-7B model by using 8 instances of Mistral-7B: for each token, the model selects two of the Mistral-7B models to produce an output, which are then combined (Jiang et al., 2024). Its performance is claimed to be equal to the Llama2-70B model (Jiang et al., 2024). In our experiments, we used the Mistral-7B model (4bit quantised to ensure consistency across all models) and the Mixtral-8x7B model (4bit quantised as the base model needs nearly 95GB of GPU space).

**OpenAI model** The **GPT-3.5-turbo** model is an LLM developed by OpenAI and has demonstrated high performance in a wide range of natural language processing tasks (Ye et al., 2023). We

chose the GPT-3.5-turbo-0125 version as it had the best performance/cost trade-off among the closed-source LLMs.

## 5 Results

### 5.1 Binary RbAM

**Baselines** Table 2 shows the results for the baselines in the binary RbAM task, i.e. RoBERTa fine-tuned on each dataset and then evaluated on the remaining datasets.

RoBERTa fine-tuned with the DP dataset achieved the highest micro $F_1$ score of 76 and an $F_1$ score better than other baselines in three datasets (Kialo, M-Arg, and ComArg). Fine-tuning took 0.23 hours for the DP dataset.

RoBERTa fine-tuned with the Kialo and the IBM datasets achieved a micro $F_1$ score of 74 and 73, respectively, which came close to the RoBERTa fine-tuned with the DP dataset. RoBERTa fine-tuned with Kialo achieved a better $F_1$ score compared to the other baselines in five datasets (NK, IBM, DP, Web, and UKP). These datasets are larger than DP and so fine-tuning took 53.73 hours for the Kialo dataset and 0.96 hours for the IBM dataset[12].

**Large Language Models** Table 3 shows the results. We can see that Mixtral-8x7B achieved the highest micro $F_1$ score of 82, outperforming all of the baselines. Also, in eight of the datasets (IBM, Essay, Kialo, M-Arg, Mic, Micro, Web, ComArg and CDCP), it achieved the highest $F_1$ score of all LLMs (as well as better than all baselines in all of these datasets except four). However, the inference time of 0.75 seconds per argument pair for this model is high compared to the baselines.

Llama2-70B performed almost as well as Mixtral-8x7B, with a micro $F_1$ score of 81. The average $F_1$ score for the support labels was the same for Llama2-70B but the average $F_1$ score for the attack labels a point lower. However, it achieved the highest $F_1$ scores in three datasets (NK, DP, and UKP). Its inference time was even a higher 1.18 seconds per argument pair.

Mistral-7B performed well given that it is smaller than the other LLMs used, achieving a micro $F_1$ score of 75, which was close to the best performing baseline. However, it did not outperform any other LLMs in any dataset. Mistral-7B was also the fastest, with an inference time of 0.19 seconds per argument pair.

[12] For all of the baseline models, a single inference took 0.005 seconds for each test sample.

GPT-3.5-turbo did not perform as well as we expected; it was the worst performing LLM in the binary RbAM task, achieving a micro $F_1$ score of 71 (which still surpassed eight of the baselines). We also tested GPT-4o-mini in a small subset of the datasets (4000 samples randomly drawn from the combination of the datasets), achieving a micro $F_1$ score of 78, i.e. better than GPT-3.5-turbo but not Mixtral-8x7B.

In conclusion, **Llama2-70B** and **Mixtral-8x7B** surpassed the baselines, including the state-of-the-art RoBERTa model, with the latter outperforming the former and also bringing the upsides of fast inference time and fewer GPU requirements.

### 5.2 Ternary RbAM

**Baseline** The left side of Table 4 shows the results for the baselines in the ternary RbAM task, i.e. RoBERTa fine-tuned on each dataset and then evaluated on the remaining datasets.

RoBERTa fine-tuned with the UKP dataset achieved the highest micro $F_1$ score of 59, outperforming other baselines in only the Web dataset. Fine-tuning took 1.95 hours for the UKP dataset.

RoBERTa fine-tuned with the NK dataset achieved a micro $F_1$ score of 57, closely matching the performance of the UKP fine-tuned model. It also surpassed other baselines in two datasets (M-Arg and UKP). The NK dataset is smaller than UKP, so its fine-tuning took only 0.28 hours[12].

**Large Language Models** The right side of Table 4 shows the results for the LLMs on the ternary RbAM task.

We can see that Mixtral-8x7B achieved the highest micro $F_1$ score of 68, outperforming all of the baselines and the other LLMs. Also, in three of the datasets (M-Arg, UKP, and NK), it achieved the highest $F_1$ score of all LLMs (as well as better than all baselines in all of these datasets except M-Arg). However, the inference time of 0.83 seconds per argument pair for this model is higher than that of the baselines.

Mistral-7B performed well given that it is smaller than the other LLMs, achieving a micro $F_1$ score of 64, which still outperformed all the baselines. However, it did not outperform the other LLMs in any dataset. Mistral-7B was also the fastest LLM, with an inference time of 0.19 seconds per argument pair.

GPT-3.5-turbo achieved a micro $F_1$ score of 60, which surpassed all of the baselines. However, it

| | RoB NK | RoB IBM | RoB Essay | RoB Kialo | RoB DP | RoB M-Arg | RoB Micro | RoB Web | RoB UKP | RoB ComArg | RoB CDCP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **NK** | - / - / - | 60 / 55 / 58 | 65 / 0 / 49 | 56 / 67 / **63** | 65 / 31 / 54 | 65 / 7 / 49 | 64 / 1 / 48 | 46 / 48 / 47 | 54 / 46 / 50 | 65 / 4 / 49 | 65 / 0 / 49 |
| **IBM** | 72 / 26 / 59 | - / - / - | 76 / 37 / 65 | 85 / 82 / **83** | 82 / 78 / 80 | 71 / 11 / 56 | 60 / 33 / 50 | 68 / 17 / 53 | 58 / 69 / 64 | 87 / 83 / 86 | 71 / 0 / 55 |
| **Essay** | 95 / 5 / 90 | 89 / 41 / 81 | - / - / - | 85 / 38 / 75 | 90 / 42 / 83 | 94 / 9 / 88 | 79 / 14 / 66 | 56 / 16 / 42 | 71 / 25 / 58 | 94 / 45 / 90 | 95 / 0 / **91** |
| **Kialo** | 68 / 14 / 53 | 74 / 73 / 73 | 70 / 18 / 56 | - / - / - | 79 / 71 / **76** | 68 / 6 / 52 | 67 / 3 / 51 | 61 / 36 / 51 | 46 / 63 / 56 | 74 / 52 / 66 | 68 / 0 / 51 |
| **DP** | 72 / 23 / 59 | 84 / 82 / 83 | 75 / 34 / 64 | 90 / 89 / **89** | - / - / - | 71 / 1 / 55 | 71 / 0 / 55 | 61 / 43 / 54 | 62 / 67 / 64 | 85 / 78 / 82 | 71 / 0 / 55 |
| **M-Arg** | 76 / 29 / 64 | 77 / 50 / 69 | 87 / 3 / 76 | 74 / 52 / 66 | 87 / 60 / **80** | - / - / - | 87 / 11 / 77 | 62 / 37 / 53 | 60 / 47 / 54 | 87 / 35 / 79 | 86 / 0 / 76 |
| **Micro** | 83 / 3 / 71 | 77 / 52 / 69 | 85 / 28 / **76** | 73 / 53 / 65 | 83 / 51 / 75 | 82 / 11 / 70 | - / - / - | 60 / 34 / 50 | 52 / 44 / 48 | 83 / 33 / 72 | 84 / 0 / 73 |
| **Web** | 67 / 15 / 52 | 65 / 60 / 63 | 68 / 13 / 53 | 67 / 67 / **67** | 65 / 59 / 62 | 61 / 43 / 54 | 61 / 40 / 53 | - / - / - | 51 / 67 / 61 | 69 / 32 / 58 | 67 / 0 / 51 |
| **UKP** | 61 / 28 / 49 | 73 / 75 / 74 | 67 / 42 / 58 | 68 / 81 / **76** | 68 / 75 / 72 | 59 / 42 / 52 | 51 / 47 / 49 | 58 / 38 / 50 | - / - / - | 74 / 67 / 71 | 61 / 0 / 44 |
| **ComArg** | 72 / 2 / 57 | 73 / 71 / 72 | 76 / 36 / 65 | 71 / 74 / 73 | 82 / 73 / **78** | 71 / 20 / 57 | 72 / 5 / 57 | 72 / 3 / 57 | 59 / 62 / 60 | - / - / - | 73 / 0 / 57 |
| **CDCP** | 98 / - / 96 | 77 / - / 63 | 100 / - / **99** | 75 / - / 60 | 90 / - / 82 | 91 / - / 83 | 95 / - / 90 | 34 / - / 20 | 42 / - / 27 | 98 / - / 96 | - / - / - |
| **Mic Avg.** | 69 / 15 / 55 | 74 / 72 / 73 | 70 / 21 / 57 | 76 / 73 / 74 | 79 / 71 / **76** | 69 / 11 / 54 | 67 / 10 / 52 | 60 / 35 / 51 | 48 / 62 / 56 | 75 / 53 / 68 | 69 / - / 52 |
| **Train T.** | 0.29hr | 0.96hr | 2.14hr | 53.73hr | 0.23hr | 0.19hr | 0.18hr | 1.07hr | 4.47hr | 0.45hr | 0.52hr |

Table 2: $F_1$ scores (as a percentage) for support / attack / both (where both is the micro average) for various datasets (rows) by the RoBERTa baselines, fine-tuned on the datasets (columns). 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.

| | Llama2-70B | Mistral-7B | Mixtral-8x7B | GPT-3.5-turbo | Avg. |
|---|---|---|---|---|---|
| **NK** | 70 / 68 / **69** | 67 / 43 / 58 | 69 / 34 / 58 | 51 / 70 / 63 | 62 |
| **IBM** | 92 / 91 / 92 | 86 / 85 / 85 | 93 / 92 / **93** | 85 / 85 / 85 | 89 |
| **Essays** | 90 / 44 / 84 | 84 / 32 / 74 | 91 / 44 / **84** | 87 / 35 / 78 | 80 |
| **Kialo** | 82 / 79 / 81 | 76 / 75 / 75 | 83 / 80 / **82** | 68 / 73 / 70 | 77 |
| **DP** | 94 / 93 / **93** | 89 / 87 / 88 | 91 / 89 / 90 | 84 / 83 / 83 | 89 |
| **M-Arg** | 77 / 60 / 71 | 72 / 57 / 66 | 81 / 60 / **75** | 75 / 56 / 68 | 70 |
| **Micro** | 78 / 48 / 69 | 75 / 47 / 66 | 81 / 50 / **72** | 71 / 46 / 62 | 67 |
| **Web** | 68 / 70 / 69 | 60 / 70 / 65 | 69 / 70 / **69** | 56 / 69 / 64 | 67 |
| **UKP** | 79 / 87 / **84** | 52 / 78 / 70 | 73 / 84 / 80 | 74 / 83 / 79 | 78 |
| **ComArg** | 75 / 73 / 74 | 69 / 71 / 70 | 80 / 76 / **78** | 58 / 70 / 65 | 72 |
| **CDCP** | 90 / - / 81 | 81 / - / 68 | 92 / - / **86** | 83 / - / 71 | 77 |
| **Mic Avg.** | 83 / 79 / 81 | 75 / 74 / 75 | 83 / 80 / **82** | 70 / 73 / 71 | 77 |
| **Inference T.** | 1.18s | 0.19s | 0.75s | 0.43s | |

Table 3: $F_1$ scores (as a percentage) for support / attack / both (where both is the micro average) relations in various datasets (rows) for the models used (columns) with 2A2S without task definition and without applying instruction template. 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.

also did not achieve better results than the other LLMs in any dataset. We also tested GPT-4o-mini in a small subset of the datasets (2000 samples randomly drawn from the combination of the datasets), achieving a micro $F_1$ score of 63, i.e. better than GPT-3.5-turbo but not Mixtral-8x7B or Mistral-7B.

Llama2-70B surprisingly, given its performance in the binary RbAM task, achieved the lowest micro $F_1$ score of 54 for the ternary task.

To conclude, **Mixtral-8x7B** and **Mistral-7B** surpassed the baselines, including the state-of-the-art RoBERTa, and the former outperformed the latter.

## 6 Ablation Studies

To validate the effectiveness of LLMs, we conducted more experiments with different variations of the main primer and prompt.

The main primer for the binary task is the 2A2S primer. The variations considered for the ablation studies were 2A1S, 1A2S, 1A1S primers, and zero-shot (without primer). For zero-shot we included the task definition for the (binary) RbAM task as this gave better performance in the initial experiments we conducted (see Appendix A). Table 5 shows the results[13] for the ablation study on the binary RbAM task. For the open-source models, the 2A2S primer performed best. For GPT-3.5-turbo, the zero-shot primer performed best, likely due to the model's ability to follow instructions (as instructions are included with zero-shot) rather than primers.

For the ternary task the main prompt is 1A1S1N primer. For the ablations we only considered the variation without the primer (zero-shot). For zero-shot we included instructions for the RbAM task as, again, this gave better results in the initial experiments we conducted (see Appendix A). Table 6 shows the results[13] for the ablation study on the ternary RbAM task. For most of the open-source models, the 1A1S1N primer performs better (except Llama2-70B, where the micro $F_1$ scores are equal). For GPT-3.5-turbo, the zero-shot primer performed slightly better.

Overall, these results show that LLMs with more (informative) examples perform better for the RbAM task. However, even with imbalanced primers or no primers, LLMs still perform close to (and sometimes better than) the baselines.

## 7 Extra Ternary RbAM Datasets

We experimented with the recent[14] ARIES benchmark (Gemechu et al., 2024) datasets, consisting of seven datasets (AMP was not available). For these datasets, we conducted experiments using the best performing RoBERTa model (fine-tuned with the UKP dataset) and the best performing LLM (the Mixtral-8x7B model with 2A2S primer)[15].

[13] For the results on individual labels, see Appendix B.

[14] The dataset was made available on 6/9/2024, which did not give us enough time to run all the experiments.

[15] We also wanted to compare LLMs with the ARIES benchmark baselines; however, they reported macro $F_1$ scores, whereas we report micro $F_1$ scores. In the future, we aim to run their baselines to compare with LLMs for RbAM.

| | Baselines | | | | LLMs for RbAM | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **RoB M-Arg** | **RoB Web** | **RoB UKP** | **RoB NK** | **Llama2-70B** | **Mistral-7B** | **Mixtral-8x7B** | **GPT-3.5-turbo** | **Avg.** |
| **M-Arg** | - / - / - / - | 22 / 10 / 81 / 69 | 19 / 4 / 91 / 83 | 3 / 6 / 93 / **87** | 30 / 11 / 66 / 51 | 27 / 15 / 80 / 66 | 19 / 19 / 89 / **80** | 33 / 16 / 78 / 65 | 66 |
| **Web** | 0 / 0 / 51 / 34 | - / - / - / - | 24 / 14 / 55 / **40** | 3 / 1 / 51 / 35 | 53 / 54 / 55 / **54** | 44 / 44 / 56 / 49 | 30 / 37 / 58 / 48 | 47 / 51 / 58 / 53 | 51 |
| **UKP** | 0 / 0 / 72 / 56 | 38 / 55 / 2 / 33 | - / - / - / - | 10 / 14 / 72 / **56** | 59 / 58 / 53 / 56 | 52 / 64 / 76 / 68 | 52 / 64 / 76 / **69** | 57 / 59 / 66 / 62 | 64 |
| **NK** | 0 / 0 / 76 / **62** | 18 / 34 / 27 / 29 | 26 / 12 / 66 / 49 | - / - / - / - | 39 / 38 / 8 / 29 | 38 / 37 / 45 / 41 | 33 / 13 / 76 / **62** | 36 / 38 / 37 / 37 | 42 |
| **Mic Avg.** | 0 / 0 / 70 / 54 | 37 / 51 / 27 / 37 | 24 / 13 / 75 / **59** | 9 / 12 / 73 / 57 | 55 / 53 / 54 / 54 | 48 / 57 / 73 / 64 | 46 / 58 / 76 / **68** | 53 / 55 / 66 / 60 | 62 |
| **Time** | 0.07hr | 0.58 | 1.95hr | 0.28hr | 1.3s | 0.19s | 0.83s | 0.43s | |

Table 4: $F_1$ scores (as a percentage) for support / attack / neither / all (where all is the micro average) relations in various datasets (rows) for the baselines (left side columns) and models (right side columns) used with 1A1S1N with task definition and without applying instruction template. RoB here stands for the RoBERTa baselines, fine-tuned on the datasets (columns). Boldface font indicates the best performing baseline and model (for all relations) for each dataset. The last row gives the time it took for fine-tuning for the baselines, in hours and the time it takes for a single inference for each LLM, in seconds.

| | **Llama2-70B** | | | | | **Mistral-7B** | | | | | **Mixtral-8x7B** | | | | | **GPT-3.5-turbo** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2A2S | 0 | 2A1S | 1A2S | 1A1S | 2A2S | 0 | 2A1S | 1A2S | 1A1S | 2A2S | 0 | 2A1S | 1A2S | 1A1S | 2A2S | 0 | 2A1S | 1A2S | 1A1S |
| **NK** | **69** | 64 | 59 | 60 | 65 | 58 | 63 | 62 | 61 | 59 | 58 | 64 | 68 | 53 | 58 | 63 | 68 | 64 | 52 | 63 |
| **IBM** | 92 | 86 | 91 | 75 | 80 | 85 | 72 | 85 | 86 | 88 | 93 | 84 | **95** | 89 | 91 | 85 | 86 | 86 | 76 | 89 |
| **Essays** | 84 | 82 | **90** | 73 | 70 | 74 | 78 | 76 | 74 | 77 | 84 | 81 | 84 | 83 | 78 | 78 | 86 | 77 | 77 | 82 |
| **Kialo** | 81 | 67 | 62 | 69 | 73 | 75 | 67 | 77 | 75 | 72 | **82** | 68 | 80 | 80 | 81 | 70 | 71 | 76 | 67 | 74 |
| **DP** | **93** | 84 | 87 | 84 | 85 | 88 | 83 | 87 | 86 | 88 | 90 | 86 | 92 | 90 | 90 | 83 | 88 | 86 | 78 | 87 |
| **M-Arg** | 71 | 73 | **78** | 40 | 51 | 66 | 71 | 55 | 55 | 68 | 75 | 70 | 71 | 70 | 73 | 68 | 75 | 68 | 64 | 67 |
| **Micro** | 69 | 63 | 67 | 55 | 59 | 66 | 62 | 63 | 66 | 71 | 72 | 70 | 70 | 72 | **73** | 62 | 70 | 61 | 62 | 67 |
| **Web** | 69 | 62 | 58 | 55 | 64 | 65 | 60 | 64 | 62 | 66 | 69 | 67 | 67 | 68 | **70** | 64 | 65 | 66 | 57 | 64 |
| **UKP** | **84** | 80 | 78 | 75 | 83 | 70 | 78 | 69 | 63 | 75 | 80 | 77 | 80 | 72 | 76 | 79 | 83 | 80 | 74 | 82 |
| **ComArg** | 74 | 73 | 77 | 64 | 70 | 70 | 72 | 75 | 65 | 70 | 78 | 73 | **82** | 72 | 76 | 65 | 79 | 70 | 46 | 64 |
| **CDCP** | 81 | 81 | **98** | 54 | 50 | 68 | 76 | 62 | 63 | 66 | 86 | 77 | 86 | 88 | 81 | 71 | 86 | 85 | 67 | 76 |
| **Mic Avg.** | 81 | 74 | 76 | 71 | 74 | 75 | 71 | 70 | 74 | 72 | **82** | 74 | 79 | 79 | 80 | 71 | 77 | 74 | 68 | 74 |

Table 5: Results from ablation studies for the binary RbAM task. Only the micro $F_1$ score has been reported for each dataset, model, primer combination. Boldface font indicates the best performing combination.

| | **Llama2-70B** | | **Mistral-7B** | | **Mixtral-8x-7B** | | **GPT-3.5-turbo** | |
|---|---|---|---|---|---|---|---|---|
| | 1A1S1N | 0 | 1A1S1N | 0 | 1A1S1N | 0 | 1A1S1N | 0 |
| **M-Arg** | 51 | 49 | 66 | 76 | **80** | 68 | 65 | 67 |
| **Web** | 54 | **58** | 49 | 43 | 48 | 57 | 53 | 52 |
| **UKP** | 56 | 58 | 68 | 59 | **69** | 59 | 62 | 65 |
| **NK** | 29 | 20 | 41 | 51 | **62** | 47 | 37 | 25 |
| **Mic Avg.** | 54 | 54 | 64 | 61 | **68** | 58 | 60 | 61 |

Table 6: Results from ablation studies for the ternary RbAM task. Only the micro $F_1$ score has been reported for each dataset, model, primer combination. Boldface font indicates the best performing combination.

| | **MTC** | **AAEC** | **CDCP** | **ACSP** | **AstRCT** | **US2016** | **QT30** |
|---|---|---|---|---|---|---|---|
| **RoB** | 45 | 46 | 51 | **66** | 45 | 62 | 64 |
| **LLMs** | **62** | **56** | **63** | 61 | **63** | **62** | **66** |

Table 7: Micro $F_1$ scores of the best baseline (RoBERTa fine-tuned with the UKP dataset) compared with the best combination of LLMs for RbAM (Mixtral-8x7B with 2A2S shot, excluding task definition and applying instruction template) on the ARIES benchmark. Boldface font indicates the best performing model.

Table 7 presents the results from the experiments on the ARIES datasets. It can be seen that Mixtral-8x7B performed better in most datasets, except ACSP. These results further indicate that LLMs are effective at RbAM.

## 8 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 eleven datasets and four LLMs, that **Llama2-70B** and **Mixtral-8x7B** surpassed the RoBERTa baseline in the binary RbAM task, with the latter outperforming the former and also having faster inference time and fewer GPU requirements. We also showed, with experiments on four datasets and four LLMs, that **Mixtral-8x7B** surpassed the RoBERTa baseline in the ternary RbAM task. Further, we demonstrated that our approach is robust to different primers with ablation studies, where the results show that more examples in the primer give better results. Finally, we showed that LLMs perform better than the RoBERTa baseline in the recently proposed ARIES benchmark.

For future work, there are many potential avenues, including the following four. 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 handle out-of-distribution scenarios as in (Waldis et al., 2024). 3) We want to test whether taking context into account when LLMs are prompted for RbAM would improve performance, as in (Mezza et al., 2024). 4) We would like to consider RbAM alongside other AM tasks, in the spirit of (Sun et al., 2024), and also leverage on insights from argumentation theory.

## 9 Limitations

There are some limitations of our work. First, even though we test LLMs for RbAM on eleven datasets, its generalisability to other domains not covered by the datasets remains uncertain. Different domains may present unique challenges where LLMs could fail. 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.

## 10 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/neither, so there are no risks of generating biased or false information.

## Acknowledgments

This research was partially supported by ERC under the EU's Horizon 2020 research and innovation programme (grant agreement No. 101020934), and by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme.

## References

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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 251–261.

Filip Boltuzic and Jan Snajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, pages 49–58.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Tweeties squabbling: Positive and negative results in applying argument mining on social media. In *Computational Models of Argument - Proceedings of COMMA 2016, Potsdam, Germany, 12-16 September, 2016*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 208–212.

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*, pages 449–450.

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, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5427–5433.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*, pages 29–34.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2309–2330.

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*, pages 45–52.

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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1374–1379.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural*

*Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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.

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.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. ARIES: A general benchmark for argument relation identification. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 1–14.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2023. Argucast: A system for online multi-forecasting with gradual argumentation. In *Proceedings of the First International Workshop on Argumentation and Applications co-located with 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023), Rhodes, Greece, September 2-8, 2023*, CEUR Workshop Proceedings, pages 40–51.

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, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 60–66.

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. *CoRR*, abs/2310.06825.

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. *CoRR*, abs/2401.04088.

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 2016, Portorož, Slovenia, May 23-28, 2016*.

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. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4889–4896.

Rafael Mestre, Razvan Milicin, Stuart Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining, ArgMining@EMNLP 2021, Punta Cana, Dominican Republic, November 10-11, 2021*, pages 78–88.

Stefano Mezza, Wayne Wobcke, and Alan Blair. 2024. Exploiting dialogue acts and context to identify argumentative relations in online debates. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 36–45.

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, May 7-12, 2018*.

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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 938–948.

Ramon Ruiz-Dolz, José Alemany, Stella Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intell. Syst.*, 36(6):62–70.

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, ArgMining 2023, Singapore, December 7, 2023*, pages 1–10.

Ameer Saadat-Yazdi and Nadin Kökciyan. 2024. Beyond recognising entailment: Formalising natural language inference from an argumentative perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9620–9636.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Comput. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 3664–3674.

Yang Sun, Muyi Wang, Jianzhu Bao, Bin Liang, Xiaoyan Zhao, Caihua Yang, Min Yang, and Ruifeng Xu. 2024. PITA: prompting task interaction for argumentation mining. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5036–5049.

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*, pages 27–44.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Dietrich Trautmann, Michael Fromm, Volker Tresp, Thomas Seidl, and Hinrich Schütze. 2020. Relational and fine-grained argument mining. *Datenbank-Spektrum*, 20(2):99–105.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaría. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining, ArgMining@COLING 2022, Online and in Gyeongju, Republic of Korea, October 12 - 17, 2022*, pages 95–103.

Andreas Waldis, Yufang Hou, and Iryna Gurevych. 2024. How to handle different types of out-of-distribution scenarios in computational argumentation? A comprehensive and fine-grained field study. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14878–14898.

Marilyn A. Walker, Jean E. 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 2012, Istanbul, Turkey, May 23-25, 2012*, pages 812–817.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *CoRR*, abs/2303.10420.

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

# Appendix

## A Preliminary Experiments

In this section, we explain the preliminary experiments that we have run to select the best primer and hyperparameters.

### A.1 2A2S and Zero-Shot for Binary RbAM

We have randomly selected four 2A2S primer seeds, consisting of two support and two attack relations from DebatepediaProcon dataset (except seed 3, which was generated by us) as the baseline trained with DP dataset has achieved the best accuracy. Then, we tested our prompt with these primer seeds on 4000 randomly selected examples from all the datasets using the Mistral-7B model. We have also tested whether including task definition and instruction template was beneficial.

| Zero-Shot | Seed 0 | Seed 1 | Seed 2 | **Seed 3** |
|---|---|---|---|---|
| 65 / 56 / 61 | 56 / 73 / 64 | 65 / 74 / 69 | 52 / 72 / 62 | 76 / 75 / **75** |
| 70 / 50 / 60 | 71 / 75 / 73 | 70 / 72 / 71 | 66 / 75 / 71 | 74 / 74 / 74 |
| 77 / 73 / **75** | 64 / 72 / 68 | 66 / 73 / 69 | 69 / 74 / 71 | 72 / 60 / 66 |
| 70 / 67 / 69 | 71 / 68 / 70 | 65 / 70 / 67 | 66 / 70 / 68 | 70 / 69 / 70 |

Table 8: Prompt selection for 2A2S: • First row **without task definition**, without applying instruction template; • second row with task definition, without applying template, • third row without task definition, with applying instruction template; • final row with task definition, with applying template.

We have decided to include task definition and not use instructions to zero-shot prompts as that combination achieved the highest accuracy. For 2A2S primer we decided not to include task definition, not to apply instruction template, and use seed 2.

### A.2 2A1S and 1A2S for Binary RbAM

We have randomly selected five 2A1S and 1A2S primer seeds, from the DebatepediaProcon dataset (except seeds 3 and 4, which was derived from the 2A2S primer). Again, we tested our prompt with these primer seeds on 4000 randomly selected examples from all the datasets using the Mistral-7B model without applying instruction template and without including task definition.

| Seed 0 | Seed 1 | **Seed 2** | Seed 3 | Seed 4 |
|---|---|---|---|---|
| 42 / 69 / 55 | 73 / 74 / 73 | 75 / 78 / **76** | 73 / 68 / 71 | 75 / 69 / 72 |

Table 9: Prompt selection for 2A1S primer.

We have found that seed 2 was better for the 2A1S primer. Therefore, seed 2 was used in the ablation studies for 2A1S primer.

| Seed 0 | Seed 1 | Seed 2 | **Seed 3** | Seed 4 |
|---|---|---|---|---|
| 61 / 73 / 67 | 69 / 76 / 73 | 68 / 73 / 70 | 74 / 74 / **74** | 66 / 74 / 70 |

Table 10: Prompt selection for 1A2S primer.

For 1A2S primer seed 3 has achieved the highest micro $F_1$ score. Therefore, seed 3 was picked for the ablation studies for 1A2S primer.

### A.3 1A1S for Binary RbAM

We have selected four 1A1S primer seeds which was derived from the 2A2S primer. Again, we tested our prompt with these primer seeds on 4000 randomly selected examples from all the datasets using the Mistral-7B model without applying instruction template and without including task definition.

| Seed 0 | **Seed 1** | Seed 2 | Seed 3 |
|---|---|---|---|
| 71 / 69 / 70 | 75 / 68 / **72** | 70 / 72 / 71 | 70 / 73 / 71 |

Table 11: Prompt selection for 1A1S primer.

Seed 1 was the best performing seed, therefore, in the ablation studies we have selected seed 1 for 1A1S primer.

### A.4 Zero-Shot and 1A1S1N for Ternary RbAM

We have selected three 1A1S1N primer seeds, consisting of one support, one attack (where these relations were derived from the 1A1S primer), and one no relations randomly drawn from the UKP dataset (as the baseline trained with UKP dataset has achieved the best accuracy). Then, we tested our prompt with these primer seeds on 2000 randomly selected examples from all the datasets using the Mistral-7B model. We have also tested whether including task definition and instruction template was beneficial.

We have decided to include task definition and use instruction template for zero-shot prompts as this combination achieved the highest accuracy. For 1A1S1N primers we decided to include task definition, not to apply instruction template, and use seed 0.

| Zero-Shot | **Seed 0** | Seed 1 | Seed 2 |
|---|---|---|---|
| 40 / 37 / 55 / 48 | 46 / 35 / 58 / 50 | 39 / 46 / 39 / 41 | 38 / 45 / 34 / 37 |
| 52 / 8 / 72 / 54 | 51 / 56 / 74 / **65** | 39 / 54 / 74 / 62 | 32 / 50 / 75 / 61 |
| 48 / 49 / 61 / 56 | 47 / 31 / 72 / 58 | 47 / 39 / 75 / 61 | 47 / 46 / 73 / 62 |
| 51 / 26 / 73 / **58** | 44 / 55 / 75 / 64 | 40 / 48 / 76 / 62 | 31 / 44 / 75 / 59 |

Table 12: Prompt selection for 1A1S1N: • First row without task definition, without applying instruction template; • second row **with task definition**, without applying template, • third row without task definition, with applying instruction template; • final row with task definition, with applying template.

## B Full Results from the Ablation Studies

This section presents the full results from the ablation studies.

### B.1 Zero-Shot Experiments

### B.2 2A1S-Shot Experiments

### B.3 1A2S-Shot Experiments

### B.4 1A1S-Shot Experiments

### B.5 Zero-Shot Ternary Experiments

## C 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.

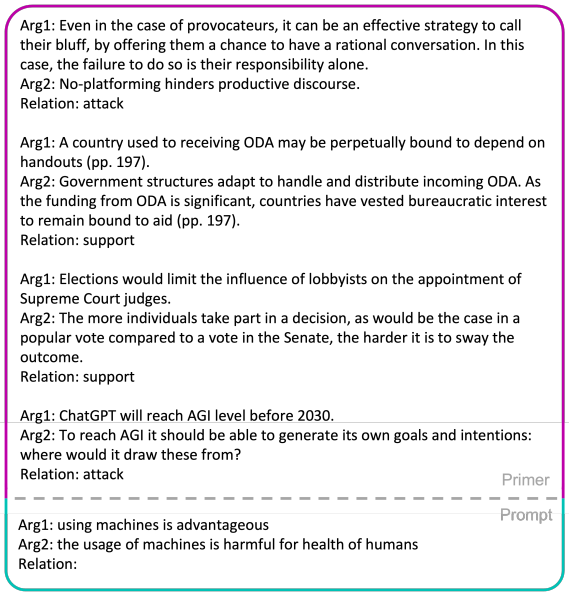


Figure 2: An example prompt drawn from the Essays dataset used in the RbAM experiments.

## D Datasets

Number of average words and characters for each dataset are given in Table 18. This kind of statistics

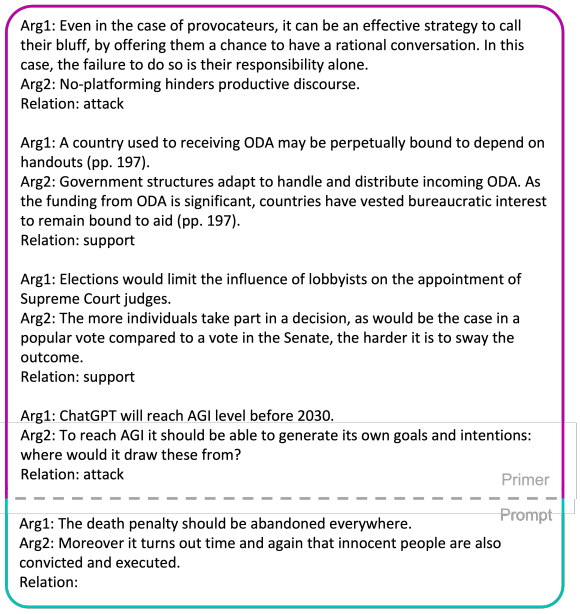


Figure 3: An example prompt drawn from the Microtexts dataset used in the RbAM experiments.

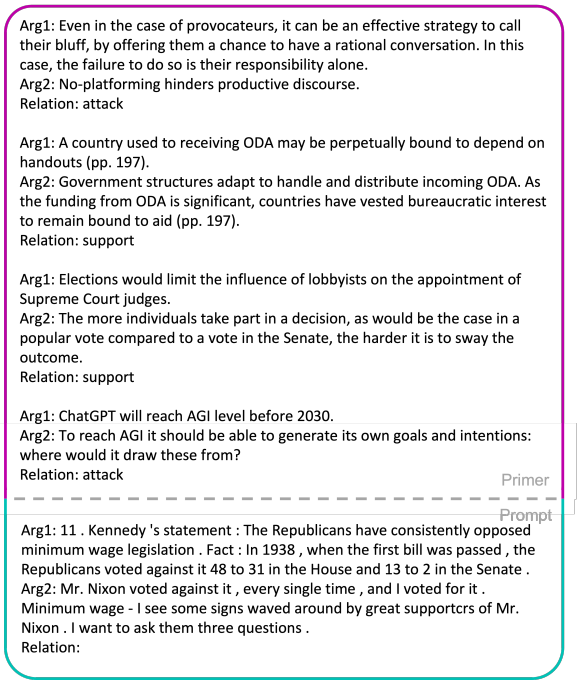


Figure 4: An example prompt drawn from the Nixon-Kennedy dataset used in the RbAM experiments.

| | Llama2-70B | Mistral-7B | Mixtral-8x7B | GPT-3.5-turbo-0125 |
|---|---|---|---|---|
| NixonKennedyDataset | 63 / 65 / 64 | 61 / 65 / 63 | 67 / 61 / 64 | 70 / 67 / 68 |
| IBMDebaterDataset | 87 / 85 / 86 | 71 / 74 / 72 | 84 / 84 / 84 | 89 / 83 / 86 |
| EssaysRelationDataset | 87 / 32 / 82 | 83 / 29 / 78 | 85 / 37 / 81 | 91 / 40 / 86 |
| KialoDataset | 71 / 63 / 67 | 68 / 66 / 67 | 73 / 61 / 68 | 77 / 65 / 71 |
| DebatepediaProconDataset | 86 / 82 / 84 | 84 / 81 / 83 | 88 / 84 / 86 | 90 / 85 / 88 |
| M-ArgDataset | 78 / 59 / 73 | 75 / 57 / 71 | 74 / 59 / 70 | 80 / 60 / 75 |
| MicrotextsDataset | 71 / 42 / 63 | 71 / 37 / 62 | 77 / 53 / 70 | 79 / 47 / 70 |
| WebDataset | 62 / 62 / 62 | 58 / 62 / 60 | 64 / 70 / 67 | 65 / 64 / 65 |
| UKPDataset | 77 / 83 / 80 | 73 / 81 / 78 | 69 / 83 / 77 | 80 / 85 / 83 |
| ComArgDataset | 74 / 70 / 73 | 71 / 74 / 72 | 72 / 74 / 73 | 81 / 76 / 79 |
| CDCPDataset | 81 / 0 / 81 | 76 / 0 / 76 | 77 / 0 / 77 | 86 / 0 / 86 |
| Avg. | 72 / 65 / 69 | 69 / 67 / 68 | 74 / 64 / 70 | 78 / 67 / 74 |

Table 13: $F_1$ scores (as a percentage) for support / attack / both (where both is the micro average) relations in various datasets (rows) for the models used (columns) with zero-shot without task definition and without applying instruction template. Boldface font indicates the best performing model (for both relations) for each dataset.

| | Llama2-70B | Mistral-7B | Mixtral-8x7B | GPT-3.5-turbo-0125 |
|---|---|---|---|---|
| NixonKennedyDataset | 68 / 50 / 59 | 53 / 70 / 62 | 72 / 64 / 68 | 61 / 68 / 64 |
| IBMDebaterDataset | 93 / 90 / 91 | 85 / 84 / 85 | 95 / 94 / 95 | 86 / 86 / 86 |
| EssaysRelationDataset | 95 / 43 / 90 | 81 / 32 / 76 | 89 / 40 / 84 | 81 / 34 / 77 |
| KialoDataset | 74 / 49 / 62 | 75 / 78 / 77 | 80 / 80 / 80 | 76 / 76 / 76 |
| DebatepediaProconDataset | 89 / 85 / 87 | 87 / 86 / 87 | 93 / 92 / 92 | 87 / 85 / 86 |
| M-ArgDataset | 85 / 56 / 78 | 57 / 50 / 55 | 75 / 58 / 71 | 72 / 55 / 68 |
| MicrotextsDataset | 84 / 20 / 67 | 69 / 46 / 63 | 76 / 54 / 70 | 66 / 50 / 61 |
| WebDataset | 70 / 46 / 58 | 56 / 71 / 64 | 64 / 69 / 67 | 64 / 69 / 66 |
| UKPDataset | 79 / 77 / 78 | 56 / 79 / 69 | 73 / 85 / 80 | 74 / 84 / 80 |
| ComArgDataset | 81 / 72 / 77 | 74 / 77 / 75 | 82 / 81 / 82 | 69 / 72 / 70 |
| CDCPDataset | 98 / 0 / 98 | 62 / 0 / 62 | 86 / 0 / 86 | 85 / 0 / 85 |
| Avg. | 76 / 53 / 68 | 74 / 77 / 76 | 80 / 80 / 80 | 76 / 76 / 76 |

Table 14: $F_1$ scores (as a percentage) for support / attack / both (where both is the micro average) relations in various datasets (rows) for the models used (columns) with 2A1S without task definition and without applying instruction template. Boldface font indicates the best performing model (for both relations) for each dataset.

| | Llama2-70B | Mistral-7B | Mixtral-8x7B | GPT-3.5-turbo-0125 |
|---|---|---|---|---|
| NixonKennedyDataset | 48 / 71 / 63 | 66 / 57 / 62 | 70 / 37 / 59 | 33 / 70 / 58 |
| IBMDebaterDataset | 72 / 78 / 75 | 86 / 85 / 86 | 89 / 88 / 89 | 73 / 79 / 76 |
| EssaysRelationDataset | 77 / 32 / 65 | 79 / 30 / 67 | 88 / 40 / 80 | 82 / 32 / 71 |
| KialoDataset | 63 / 76 / 71 | 76 / 74 / 75 | 81 / 78 / 80 | 60 / 73 / 68 |
| DebatepediaProconDataset | 83 / 85 / 84 | 87 / 85 / 86 | 91 / 89 / 90 | 77 / 81 / 79 |
| MArgDataset | 39 / 45 / 42 | 56 / 50 / 53 | 74 / 58 / 68 | 66 / 54 / 61 |
| MicrotextsDataset | 57 / 51 / 54 | 72 / 51 / 64 | 79 / 53 / 71 | 66 / 50 / 60 |
| WebDataset | 40 / 71 / 61 | 54 / 71 / 64 | 65 / 71 / 68 | 45 / 70 / 61 |
| UKPDataset | 66 / 83 / 77 | 46 / 77 / 68 | 61 / 81 / 74 | 65 / 82 / 76 |
| ComArgDataset | 59 / 71 / 66 | 63 / 69 / 66 | 72 / 73 / 72 | 32 / 65 / 54 |
| CDCPDataset | 54 / 0 / 37 | 63 / 0 / 46 | 88 / 0 / 79 | 67 / 0 / 50 |
| Mic Avg. | 63 / 76 / 71 | 74 / 73 / 74 | 80 / 78 / 79 | 62 / 73 / 68 |

Table 15: $F_1$ scores (as a percentage) for support / attack / both (where both is the micro average) relations in various datasets (rows) for the models used (columns) with 1A2S without task definition and without applying instruction template. Boldface font indicates the best performing model (for both relations) for each dataset.

| | Llama2-70B | Mistral-7B | Mixtral-8x7B | GPT-3.5-turbo-0125 |
|---|---|---|---|---|
| NixonKennedyDataset | 57 / 71 / 65 | 68 / 45 / 59 | 69 / 34 / 58 | 51 / 70 / 63 |
| IBMDebaterDataset | 78 / 81 / 80 | 89 / 86 / 88 | 92 / 91 / 91 | 89 / 88 / 89 |
| EssaysRelationDataset | 81 / 33 / 70 | 86 / 30 / 77 | 87 / 39 / 78 | 89 / 35 / 82 |
| KialoDataset | 68 / 77 / 73 | 75 / 67 / 72 | 82 / 80 / 81 | 74 / 73 / 74 |
| DebatepediaProconDataset | 84 / 85 / 85 | 89 / 86 / 88 | 91 / 89 / 90 | 88 / 86 / 87 |
| MArgDataset | 53 / 49 / 51 | 74 / 57 / 68 | 80 / 61 / 73 | 74 / 56 / 67 |
| MicrotextsDataset | 65 / 50 / 59 | 79 / 51 / 71 | 81 / 56 / 73 | 77 / 44 / 67 |
| WebDataset | 49 / 72 / 64 | 64 / 67 / 66 | 68 / 71 / 70 | 59 / 68 / 64 |
| UKPDataset | 78 / 87 / 83 | 68 / 80 / 75 | 64 / 82 / 76 | 78 / 84 / 82 |
| ComArgDataset | 67 / 72 / 70 | 72 / 68 / 70 | 78 / 74 / 76 | 54 / 70 / 64 |
| CDCPDataset | 66 / 0 / 50 | 79 / 0 / 66 | 90 / 0 / 81 | 87 / 0 / 76 |
| Mic Avg. | 70 / 77 / 74 | 76 / 68 / 72 | 81 / 79 / 80 | 75 / 74 / 74 |

Table 16: $F_1$ scores (as a percentage) for support / attack / both (where both is the micro average) relations in various datasets (rows) for the models used (columns) with 1A1S without task definition and without applying instruction template. Boldface font indicates the best performing model (for both relations) for each dataset.

| | Llama2-70B | Mistral-7B | Mixtral-8x7B | GPT-3.5-turbo-0125 |
|---|---|---|---|---|
| M-ArgDataset | 27 / 13 / 52 / 49 | 30 / 18 / 82 / **76** | 30 / 14 / 74 / 68 | 28 / 21 / 72 / 67 |
| WebDataset | 57 / 49 / 68 / **58** | 48 / 21 / 60 / 43 | 56 / 47 / 67 / 57 | 55 / 39 / 61 / 52 |
| UKPDataset | 52 / 55 / 60 / 58 | 51 / 33 / 73 / 59 | 54 / 57 / 62 / 59 | 56 / 60 / 71 / **65** |
| NixonKennedyDataset | 38 / 35 / 10 / 20 | 41 / 20 / 64 / **51** | 42 / 18 / 58 / 47 | 35 / 22 / 23 / 25 |
| Mic Avg. | 50 / 49 / 58 / 54 | 29 / 48 / 73 / 61 | 53 / 51 / 65 / 58 | 54 / 51 / 69 / **61** |

Table 17: $F_1$ scores (as a percentage) for support / attack / neither / both (where both is the micro average) relations in various datasets (rows) for the models used (columns) with zero-shot with task definition and with applying instruction template. Boldface font indicates the best performing model (for both relations) for each dataset.

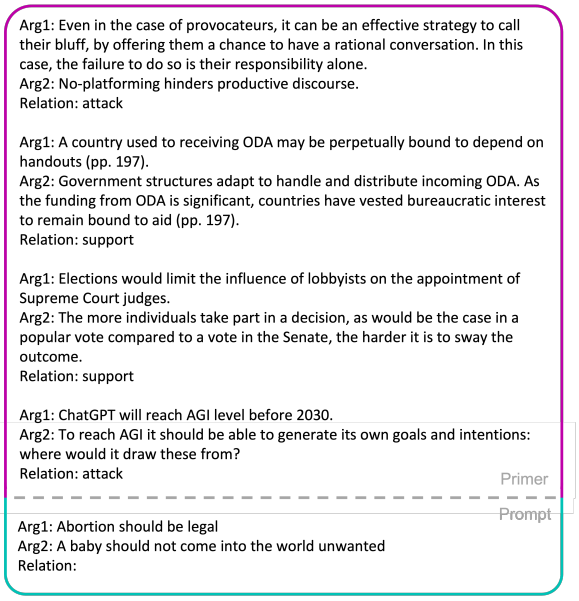


Figure 5: An example prompt drawn from the Debatepedia/Procon dataset used in the RbAM experiments.

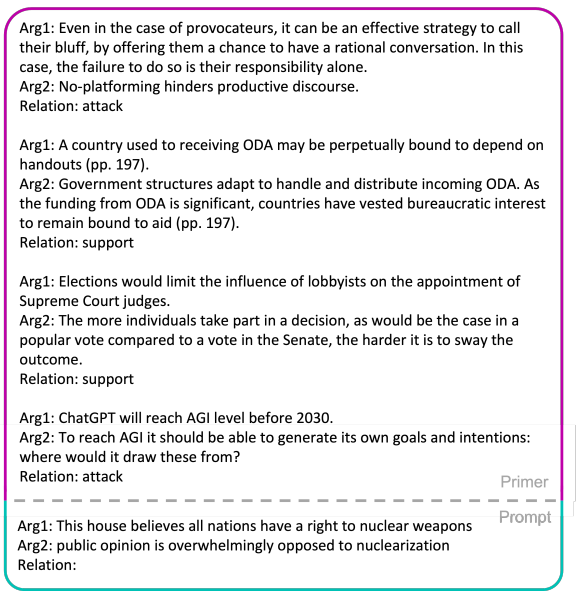


Figure 6: An example prompt drawn from the IBM-Debater dataset used in the RbAM experiments.

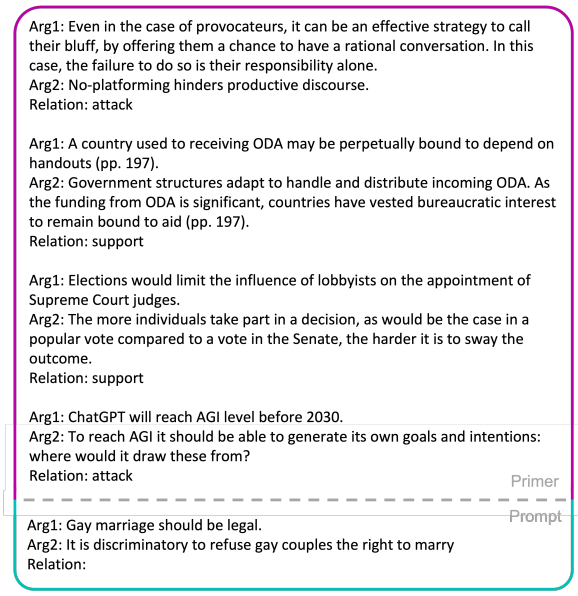


Figure 7: An example prompt drawn from the ComArg dataset used in the RbAM experiments.

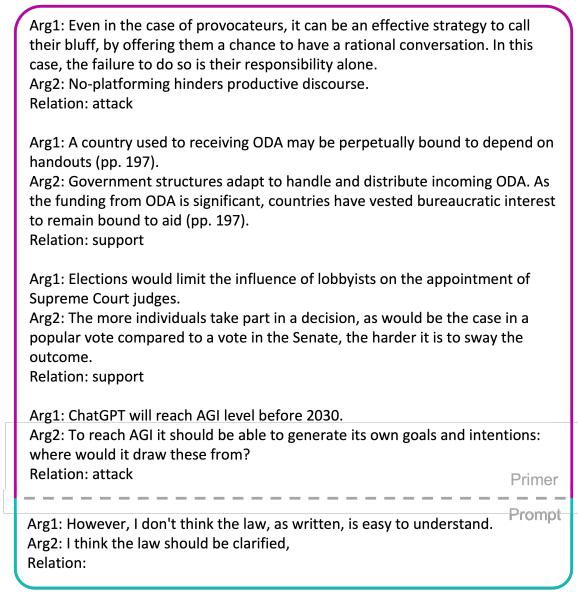


Figure 8: An example prompt drawn from the CDCP dataset used in the RbAM experiments.

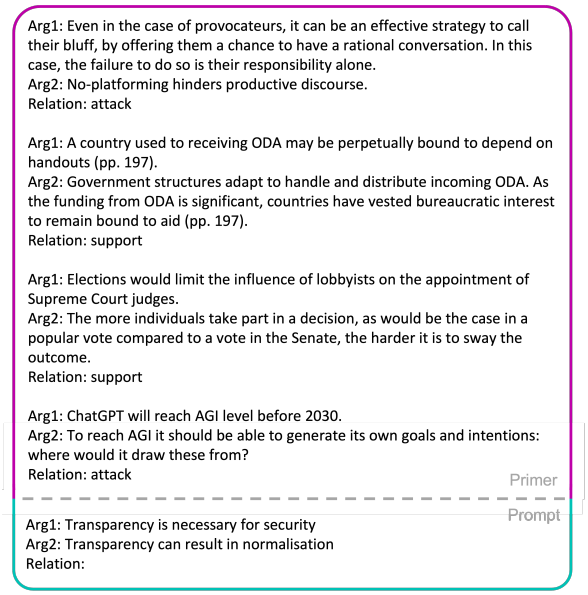


Figure 9: An example prompt drawn from the Web-Content dataset used in the RbAM experiments.

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 18: Statistical features of each dataset.