Large Language Models in Argument Mining: A Systematic Review of Performance across Tasks with Topical Arguments

Anonymous ACL submission

Abstract

Automated large-scale analysis of online argumentation around contested issues like abortion requires to detect and understand the usage of recurring arguments. Despite a large body of work in computational argumentation analysis, these tasks have not been systematically tested with large language models. We fill this gap using a data set of over 2,000 opinion comments on polarizing topics and define three tasks: argument detection, extraction and understanding whether arguments are supported or attacked in a comment. We systematically compare the performance of three state-of-the-art large language models (LLMs) and a fine-tuned RoBERTa baseline. We find that while LLMs outperform the RoBERTa baseline on argument detection and relationship classification, the simple RoBERTa model performs competitively with the LLMs on the argument extraction task, and that performance does not consistently improve with one-shot and few-shot approaches. We conclude by discussing the implications and limitations of current LLMs in argument-based opinion mining.

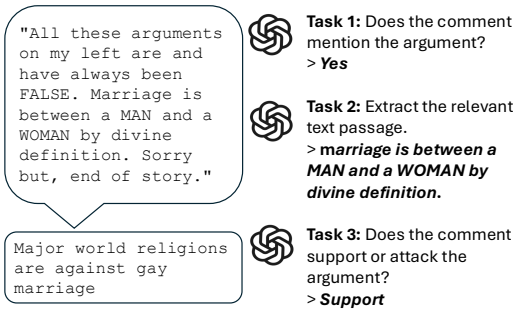

Figure 1: An opinion piece (top, left) and pre-defined argument (bottom, left). We predict whether the opinion makes use of the argument (Task 1), whether it supports or attacks the argument (Task 2) and where it mentions the argument (Task 3).

1 Introduction

Argumentation is the study of how humans express opinions, persuade others and reach conclusions, and is fundamental to human discourse and reasoning. In both formal and informal contexts, arguments form the basis of rational discourse, allowing individuals to present their viewpoints, support them with evidence, and engage in meaningful dialogue. The analysis of argumentative discourse has become increasingly critical in the digital age, where an unprecedented volume and velocity of online discourse shapes public opinion, policy decisions, and social movements (Lippi and Torroni, 2016).

This explosion of online discourse brings both challenges and opportunities for understanding human reasoning and opinion formation at scale. Automatic analysis of argumentative structures is crucial for tracking how opinions form and spread, identifying the evidence supporting different viewpoints, and evaluating the quality of public discourse (Stede and Schneider, 2018).

Public discourse around contested issues — from abortion over immigration to climate change — is often dominated by recurring arguments that are repeated over and over by different parties. To automatically 1) identify these arguments; 2) detect them in the discourse and 3) understand how they are used (supported or attacked) would be an important contribution to automatic argument analysis. However, the majority of methods in opinion mining (Sun et al., 2017; Lawrence and Reed, 2015) and sentiment analysis (Bakliwal et al., 2013; Elghazaly et al., 2016; Ramteke et al., 2016) fall short of 2) and 3) by not identifying individual arguments and their way of use. Most work in argument mining, on the other hand, focuses on individual premises and claims without abstracting to broader cross-cutting arguments (Habernal and Gurevych,

2017; Lawrence and Reed, 2019). This paper fills this gap by systematically testing state-of-the-art LLMs on a variety of tasks that would bring us closer to this goal.

Concretely, we here focus on the detection and way of usage of a set of pre-defined arguments in tasks 2, 3 – because we have to ensure that these tasks succeed before we can develop methods that identify arguments from the bottom up (task 1). We leverage data sets comprising over 2,000 opinion comments covering six polarizing topics, from gay marriage to marijuana legalization (Boltužić and Šnajder, 2014; Hasan and Ng, 2014). Each topic is associated with a pre-defined set of established arguments, and each comment was manually annotated for the presence of each argument, and its usage (support or attack). Figure 1 (left) shows an example comment-argument pair, and Figure 2 shows excerpts of comments that support or attack an argument.

Given a pair of an opinionated text and an argument, we decompose our objective into three tasks (Figure 1, right) given a comment-argument pair: 1) predict whether the comment mentions the argument; 2) extract the span that expresses the argument; 3) identify whether the argument is *supported* or *attacked* in the comment.

We experiment with three state-of-the-art large language models (GPT4o-mini, Gemini1.5-flash and Llama3-8b). We find that LLMs outperform the RoBERTa baseline on argument detection (marginally) and on relationship identification (consistently), but that the simple fine-tuned RoBERTa baseline performs at least as well as the LLMs on identifying the relevant spans of text that use an argument. Additionally, we find that using one example in the prompt (one-shot approach) generally increases performance, but that it is not necessarily the case with more examples (few-shot approach). We conclude with a discussion around implications, and important directions for future work to support accurate and reliable large-scale analysis of public discourse.

In sum, the contributions of this paper are:

- A novel lens on the problem of argument mining with a focus on identifying the presence and use of major cross-cutting arguments
- A systematic validation of three state-of-the-art LLMs across three tasks, showing that increasing instruction examples does not always improve performance and that smaller, task-specific models can compete with larger LLMs
- A discussion about implications of our findings and pointers to future work

2 Related Work

Argument mining A vast body of work exists on the mechanisms of argumentation, both from theoretical and empirical points of view. Conceptually, argument structure analysis departs from the identification of key argumentative elements, most typically classified as premises and claims (Habernal and Gurevych, 2017; Hidey et al., 2017; Feng and Hirst, 2011). Claims represent the speaker's stance or position on a topic, while premises provide supporting evidence or justification for these claims (Hidey et al., 2017; Palau and Moens, 2009).

A second major task involves determining relationships between argumentative components by analyzing how different arguments interact with each other, with the primary goal to recognize whether a premise *attacks* or *supports* a claim (Cocarascu and Toni, 2017; Carstens and Toni, 2015; Ruiz-Dolz et al., 2021; Bench-Capon, 2003). Often argument identification and relation classification are performed jointly (Egawa et al., 2020; Stab and Gurevych, 2017).

Argument structure analysis faces significant challenges, as the identification of claims remains highly subjective, with no clear linguistic consensus on their precise definition or characteristics (Daxenberger et al., 2017), and is correspondingly hard to evaluate (Mestre et al., 2022). Furthermore, most work identifies arguments on an ad-hoc, document-level basis without mapping them back to broader recurring claim types which cross-cut the discourse making them less useful to map out patterns in broader discourse. We fill this gap by testing LLMs for identifying cross-cutting arguments and their relations and use predefined premises (topical arguments) in this study to circumvent the challenge of evaluating model-identified claims.

Argument-based opinion analysis combines stance detection with argument structure into a framework for analyzing how people express their views (Arumugam, 2022), moving beyond sentiment or stance classification which typically attach binary labels, disregarding the structure of reasoning underlying the opinion piece (Hogenboom

et al., 2010).

We build on early work in this direction, which developed specialized corpora for studying argumentation by intersecting opinion-rich texts on a divisive issue (like abortion) with a pre-defined list of issue-related arguments from online debate platforms (Boltužić and Šnajder, 2014; Hasan and Ng, 2014). In Boltužić and Šnajder (2014), each comment-argument pair was manually labeled for the relationship of the argument and the comment (support or attack) and SVMs were trained to predict this relationship. Hasan and Ng (2014) retrieved list of pro and con arguments on four subtopics and manually annotated, for each comment and associated arguments, the relevant spans of the comment that expresses the argument. They trained a Maximum Entropy model with dependency-based features and joint learning to determine the reasons (arguments) on both the post and sentence level. Similarly, Sobhani et al. (2015) implemented a predefined argument approach for document-level stance classification, employing topic modeling techniques like non-negative matrix factorization. In their analysis, the authors collected and annotated news comments related to a health study leveraging arguments derived from the topic model. Here, for the first time, we systematically review the ability of state-of-the-art LLMs on these tasks.

Argument mining with Large Language Models Recently, LLMs have shown impressive performance in a variety of natural language tasks (Raiaan et al., 2024; Karanikolas et al., 2023), and argument mining is no exception. Recent works on argument pair extraction (de Wynter and Yuan, 2024), relation-based argument mining (Gorur et al., 2024; Otiefy and Alhamzeh, 2024), argument quality prediction (van der Meer et al., 2022) have shown performance gains with state-of-the-art LLMs. However, some other works have highlighted limited performance of LLMs in argumentation tasks, in particular in argument generation and persuasiveness (Hinton and Wagemans, 2023) and the identification of argumentative fallacies (Ruiz-Dolz and Lawrence, 2023).

The most comprehensive systematic review of LLMs performance in argument mining (AM) and argument generation tasks to date is Chen et al. (2024). The authors performed zero-shot and k-shot experiments using GPT-3.5, Flan-T5 family models as well as Llama2 models on a variety of AM tasks (claim detection, evidence detection, stance detection, evidence classification), as well as argument generation and summarization. Their results highlight decent performance on binary classification tasks (GPT3-5-Turbo and Flan-UL2 performing best in the zero-shot setup), but worse with more complex, multi-label classification tasks. Interestingly, a separate review focussed on the legal domain by Alsubhi et al. (2023) (including GPT-3.5 and GPT-4) concluded that these LLMs did not surpass their baseline (domain-specific BERT model). The same models also did not surpass the RoBERTa baseline in Ruiz-Dolz et al. (2024).

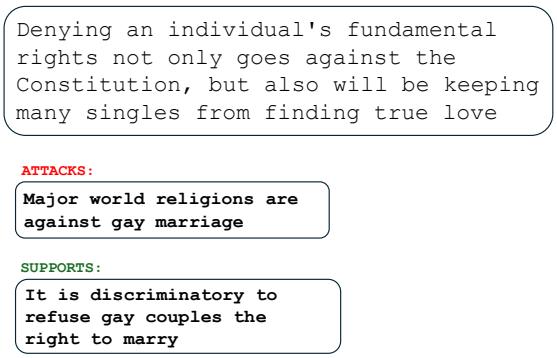

Figure 2: Illustration of a comment supporting a pro-same-sex marriage argument and attacking a con-same-sex marriage argument.

Our work complements the above in several ways. First, the capabilities of LLMs to identify cross-cutting arguments within the context of a polarizing issue have not yet been investigated. Given the contradicting results from previous surveys, we provide an important additional case study. By shedding light on three well-defined tasks we are able to identify concrete shortcomings and formulate recommendations for future work. Assessing the abilities of the LLMs to perform these tasks is of crucial importance to analyze opinions not just on a general level (such as stance, or the argumentation structure), but to support a fine-grained nuanced analysis of the arguments on which such stance is based, allowing for greater scalability across topics.

3 Methodology

3.1 Data Sets

Our study builds on prior research in natural language processing, particularly works that intersected curated arguments from online debate platforms with large-scale online argumentation data.

The **COMARG data set**: Boltužić and Šnajder (2014) manually annotated 373 comments from the discussion platform *Procon.org* with a pre-defined list of arguments retrieved from *Idebate.org*. It encompasses two topics: gay marriage and the inclusion of the phrase "Under God" in the U.S. Pledge of Allegiance. The gay marriage-related comments were annotated for three arguments in favor and four arguments against the topic, while the Pledge of Allegiance topic featured three pro and three against arguments. Each comment-argument pair was further classified based on whether the comment supported, attacked, or made no use of the argument, as well as whether the support/attack was explicit or implicit. The inter-annotator agreement was moderate, and the final labels were decided by majority vote, excluding comment-argument pairs where no majority was reached.

The **YRU data set**: Hasan and Ng (2014) sourced 1900 comments from an online debate platform, and their data set spans four topics: abortion, gay rights, legalization of marijuana, and the Obama presidency. For each topic, annotators identified a set of recurring arguments in the topics, leading to between 6 and 9 arguments each supporting and opposing the topic. The data set was originally developed for the task of argument extraction, i.e., identifying spans of text that employed a specific argument. Annotator agreement on this labelling task was reported as moderate to high, and disagreements were resolved through discussion.

All arguments for all six topics across the two data sets are listed in Table 4 in the Appendix.

3.2 Task Definitions

We define three argument mining (AM) tasks designed to test models' abilities to *detect* and *understand the use of* recurring arguments in collections of opinion texts.

Binary Argument Detection Given an argument A and a comment C, the task is to classify, in binary fashion, whether C makes use of A. We run this task on both the YRU anc COMARG data, across a total of six topics.

Argument Extraction Given an argument A and a comment C, the goal is to automatically detect the span within C that expresses A. Only the COMARG data set comes with manually-annotated argument spans, so we evaluate this task over the four COMARG topics.

Argument Relationship Classification Given an argument A and a comment C, we determine the relationship between A and C as C either attacking or supporting A (cf., Figure 2). We consider two formulations of this task: either a binary classification as support or attack; or a 4-way classification distinguishing between explicit/implicit support for or an explicit/implicit attack of an argument. Only the YRU data set labels the type of usage of an argument, so we evaluate relation classification over the two topics in this data set. A list of all pre-defined arguments from the original dataset is reported in the Appendix (Table 4).

3.3 Data Pre-Processing

Binary argument detection was conceptualized as a binary classification problem, necessitating preprocessing of the original data sets to conform to a binary format. For the COMARG data set we consider all comment-argument pairs labeled as exhibiting any form of argumentative relationship as present (1). The data contained an explicit label of 'makes no use of an argument', which we reuse as our negative (not present) label (0). The YRU data set is annotated for arguments on the sentence level. We project these labels to the comment-level, and consider them as present (1). All arguments not identified in any sentence were labeled as not present (0).

For Task 2, we treated the data in the COMARG data set differently for the two subtasks. In subtask 2a we conflated the original labels in a binary fashion, only aiming at identifying whether the comment supports or attacks the argument. For subtask 2b we considered the original scale of implicit/explicit support and attack, we thus left the original labels unaltered. Finally, for Task 3 we only considered the labels present in the original YRU data set and the manually annotated spans in the comment.

3.4 Models

We selected three Large Language Models (LLMs) from different model families, spanning both open-source and proprietary architectures: one open-source – Llama3-8b-Instruct (Dubey et al., 2024) – and two proprietary models; GPT4o-mini (Achiam et al., 2023) and Gemini1.5-Flash (Reid et al., 2024). To minimize non-deterministic behavior in the models' outputs, we followed established practices for controlling generation parameters (Zhang et al., 2023; Meng et al., 2023). Specifically, we set

Model	GM			UG			AB			GR			MA			OBI			Combined		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority	0.33	0.50	0.40	0.34	0.50	0.41	0.45	0.50	0.47	0.44	0.50	0.47	0.43	0.50	0.46	0.46	0.50	0.48	0.40	0.50	0.44
RoBERTa																			0.67	0.60	0.61
	Zero shot																				
Llama3-8b	0.67	0.66	0.66	0.64	0.66	0.65	0.64	0.74	0.69	0.63	0.71	0.67	0.62	0.68	0.65	0.59	0.69	0.64	0.63	0.69	0.65
GPT4o-m	0.74	0.70	0.72	0.67	0.67	0.67	0.69	0.83	**0.72**	0.63	0.72	0.64	0.63	0.70	0.65	0.61	0.76	0.63	0.65	0.74	0.69
Gemini1.5-f	0.77	0.78	**0.77**	0.67	0.70	0.69	0.63	0.79	0.62	0.62	0.72	0.62	0.63	0.69	0.64	0.60	0.72	0.62	0.63	0.74	0.68
	One shot																				
Llama3-8b	0.64	0.64	0.64	0.63	0.64	0.63	0.61	0.74	0.67	0.63	0.72	**0.67**	0.57	0.64	0.60	0.58	0.71	0. 64	0.61	0.68	0.64
GPT4o-m	0.75	0.70	0.72	0.64	0.65	0.65	0.68	0.84	**0.72**	0.63	0.72	0.64	0.66	0.71	**0.67**	0.63	0.70	**0.65**	0.66	0.73	0.69
Gemini1.5-f	0.70	0.71	0.70	0.66	0.67	0.66	0.62	0.80	0.62	0.63	0.74	0.64	0.59	0.68	0.59	0.59	0.71	0.60	0.63	0.74	0.68
	Five shot																				
Llama3-8b	0.62	0.64	0.63	0.59	0.60	0.60	0.57	0.68	0.62	0.60	0.71	0.65	0.59	0.70	0.64	0.55	0.68	0.61	0.59	0.67	0.63
GPT4o-m	0.69	0.67	0.68	0.67	0.67	0.67	0.66	0.81	0.68	0.61	0.72	0.61	0.64	0.70	0.66	0.60	0.75	0.61	0.63	0.73	0.68
Gemini1.5-f	0.71	0.73	0.72	0.75	0.79	**0.77**	0.67	0.82	0.70	0.64	0.74	0.65	0.60	0.71	0.58	0.61	0.72	0.63	0.65	0.77	**0.70**

Table 1: Results for binary argument detection (Task 1) across six contentious topics and the combined data set (final column) as macro-averaged precision, recall and F1. We report a majority baseline, RoBERTa (combined data only) and our three LLMs. The best F1 scores per data set are bolded.

the temperature parameter to 0 and the top_p parameter to 1, as these settings have been shown to reduce output variability and increase reproducibility (Liu et al., 2023; Brown et al., 2023). [1]

We formulate each task in a 0-shot, 1-shot and 5-shot setting. For all three tasks, we manually selected one example (for 1-shot) and five examples (for 5-shot) from each subtopic data set, ensuring a good coverage of classes and arguments in the five-shot case. The prompts used for each task can be found in the Appendix B (Tables 5 to 8).

3.5 Baseline

We fine-tuned a RoBERTa model (Liu, 2019) for each of our three tasks. Due to RoBERTa's supervised nature with high data-demand we trained only one RoBERTa model per task, combining the labelled data from all available topics.

For the classification tasks, we concatenated each comment-argument pair using the [SEP] token as a delimiter. We randomly split the data into five stratified folds for cross-validation, ensuring a balanced label distribution in each split. Each model was trained for 3 epochs with a batch size of 16. For the span extraction task, we formatted the data equivalent to extractive question-answer tasks, where arguments serve as "question", and relevant spans as the "answer" to be extracted. We fine-tuned a RoBERTa model on this data using the QuestionAnsweringModel from SimpleTransformers[2] again with five fold stratified cross validation, training for a total of 10 epochs and with a batch size of 16.

4 Results

We now present the quantitative results of our three LLMs and baselines across tasks. Overall, we find that (1) Fine-tuned RoBERTa can detect and extract arguments from comments with an accuracy close to the one observed for the LLMs; (2) that instruction examples (one- or five-shot) do not necessarily lead to enhanced performance; and (3) that the *identification* of arguments in comments (Task 1) is challenging for LLMs, which calls for caution with and future research on automated argument extraction.

4.1 Task 1: Binary Argument Detection

We test three models (Llama, GPT, Gemini) in zero-, one-, and five-shot settings across six different topics on predicting in a binary fashion whether a given argument is stated in a comment or not. Results in Table 1 show that overall GPT and Gemini outperform Llama, with all models clearly outperforming a majority baseline. The performance varies across topics, with gay marriage (GM), under god in pledge (UG) and abortion (AB) yielding the highest F1 scores.

[1] For Llama3-8b-Instruct, we also set the top_k parameter to 1. GPT40-mini and Gemini1.5Flash do not feature manual configuration of this parameter.

[2] https://simpletransformers.ai/docs/qa-model/

Model	AB			GR			MA			OB			Combined		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
RoBERTa													0.45	0.41	**0.44**
	Zero shot														
Llama3-8b	0.36	0.34	0.36	0.33	0.31	0.33	0.32	0.31	0.32	0.34	0.33	0.34	0.33	0.32	0.33
GPT4o-m	0.30	0.29	0.30	0.31	0.29	0.31	0.27	0.26	0.27	0.28	0.26	0.28	0.29	0.27	0.29
Gemini1.5-flash	0.37	0.36	0.37	0.38	0.37	0.38	0.36	0.35	0.36	0.34	0.33	0.34	0.36	0.35	0.36
	One shot														
Llama3-8b	0.33	0.31	0.33	0.38	0.37	0.38	0.31	0.28	0.30	0.36	0.35	0.36	0.34	0.32	0.34
GPT4o-m	0.31	0.30	0.32	0.31	0.29	0.31	0.33	0.31	0.33	0.33	0.31	0.33	0.32	0.30	0.32
Gemini1.5-flash	0.37	0.36	0.37	0.41	0.40	**0.41**	0.40	0.38	0.40	0.41	0.40	0.41	0.39	0.38	0.39
	Five shot														
Llama3-8b	0.38	0.37	0.38	0.39	0.37	0.39	0.36	0.35	0.36	0.37	0.35	0.37	0.37	0.36	0.37
GPT4o-m	0.32	0.30	0.31	0.36	0.35	0.36	0.37	0.35	0.37	0.37	0.35	0.37	0.35	0.33	0.35
Gemini1.5-flash	0.45	0.42	**0.43**	0.41	0.40	**0.41**	0.46	0.45	**0.46**	0.48	0.47	**0.48**	0.45	0.43	**0.44**

Table 2: Results for Argument Extraction (Task 2) for the four topics in the YRU data set and the combined data set (final column) as Rouge 1, 2 and L. We compare a fine-tuned RoBERTa model (combined data only) and our three LLMs. The best Rouge-L scores per data set are bolded.

LLMs generally outperform the established RoBERTa baseline, although mostly only marginally. In addition, while we do see a general performance boost when moving from zero-shot to one-shot settings, additional training examples (five-shot) do not bring additional advantage. This is the case across all three models.

4.2 Task 2: Argument Extraction

Here, we task models with identifying the exact span of text in which an argument is being used. We evaluated this task using ROUGE-1, -2, and-L (Lin, 2004) measuring the n-gram overlap between predicted and golden spans. Results in Table 2 reveal that, as in previous tasks, Gemini consistently demonstrates the best performance, and we see a consistent advantage of exposure to examples in the task instruction. Gemini achieves the highest scores in the five-shot settings across all topics, while GPT generally performs second-best, showing improvement with additional shots, though not as pronounced as Gemini. Its performance is relatively consistent across topics. As in task 1, we observe a very strong performance of the supervised, fine-tuned RoBERTa model and only Gemini (barely) reaches comparable performance when instantiated with five examples.

4.3 Task 3: Argument Relationship Classification

Given a comment and an argument featured in the comment, we ask models whether the argument is *supported* or *attacked* in the argument, either in a **binary** fashion, or through a 4-way **scale** (strongly/weakly supports; weakly/strongly supports). Results in Table 3 highlight notable differences in performance across models, tasks, and shot settings. While all models outperform the majority and supervised RoBERTa baselines, Gemini consistently emerges as the best-performing model across both binary and full-scale classifications, and across both topics (GM and UG). In the binary task, Gemini achieves impressive F1 scores ranging from 0.71 to 0.88, significantly outperforming both GPT and Llama. This performance gap is maintained in the more challenging full-scale task, where Gemini's F1 scores (0.38 to 0.49) still lead, albeit with a smaller margin. Across all models, performance on the binary task is substantially better than on the full-scale task, highlighting the increased complexity of distinguishing between explicit and implicit support/attack relationships.

Interestingly, performance across models was higher in the binary version of Task 3 than Task 1 for the corresponding best-performing model in each task (respectively Gemini1.5-flash 0-shot and Gemini1.5-flash 5-shot). This highlights that the models do better at identifying whether a comment supports or attacks a given argument, rather than detecting whether a comment uses the argument. Similar to the findings in Task 1 and contrary to Task 2, the impact of increasing instruction examples (from zero to five-shot) is inconsistent across models and tasks. For Gemini, performance peaks

Model	GM - binary			UG- binary			Comb- binary			GM - scale			UG - scale			Comb - scale		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority	0.31	0.50	0.39	0.29	0.50	0.37	0.30	0.50	0.38	0.10	0.25	0.14	0.29	0.50	0.37	0.19	0.37	0.25
RoBERTa							0.31	0.50	0.39							0.22	0.25	0.15
	Zero shot																	
Llama3-8b	0.40	0.40	0.40	0.64	0.60	0.62	0.52	0.50	0.51	0.24	0.25	0.24	0.39	0.42	0.40	0.31	0.33	0.31
GPT4o-m	0.50	0.50	0.50	0.77	0.78	0.78	0.63	0.64	0.64	0.24	0.23	0.24	0.23	0.42	0.29	0.23	0.33	0.26
Gemini1.5-flash	0.81	0.83	**0.82**	0.87	0.88	**0.88**	0.84	0.85	**0.85**	0.46	0.43	**0.44**	0.36	0.42	0.39	0.41	0.42	0.41
	One shot																	
Llama3-8b	0.63	0.61	0.62	0.73	0.74	0.74	0.68	0.67	0.68	0.25	0.27	0.26	0.38	0.37	0.37	0.31	0.32	0.31
GPT4o-m	0.43	0.43	0.43	0.78	0.78	0.78	0.60	0.60	0.60	0.22	0.22	0.22	0.31	0.39	0.34	0.26	0.30	0.28
Gemini1.5-flash	0.80	0.82	0.81	0.82	0.83	0.82	0.81	0.82	0.81	0.43	0.42	0.42	0.54	0.45	**0.49**	0.48	0.43	**0.45**
	Five shot																	
Llama3-8b	0.72	0.70	0.71	0.68	0.65	0.67	0.70	0.67	0.69	0.33	0.30	0.32	0.37	0.31	0.33	0.35	0.30	0.32
GPT4o-m	0.55	0.55	0.55	0.76	0.72	0.74	0.65	0.63	0.64	0.24	0.25	0.25	0.28	0.35	0.31	0.26	0.30	0.28
Gemini1.5-flash	0.77	0.76	0.76	0.74	0.68	0.71	0.75	0.72	0.73	0.39	0.43	0.41	0.41	0.36	0.38	0.40	0.39	0.39

Table 3: Results for Argument Relationship Classification (Task 3) for the two topics in the COMARG data set and the combined data set (final column) as macro precision, recall and F1. Left: binary classification (support vs attack); Right: 4-way classification (explicit/implicit support/attack). We compare against a majority baseline and fine-tuned RoBERTa model (combined data only). The best F1 scores per data set are bolded.

in the zero-shot or one-shot settings, while Llama tends to benefit from instruction examples.

4.4 Exploratory Analysis

A natural question following from the results above, is *why* LLMs fail on fine-grained argument detection and interpretation. As a step towards answering this question we conducted an exploratory analysis on the argument detection (Task 1), which is the most comprehensive in terms of samples, and which revealed substantial room for improvement. We inspected the results in Table 1 separately by argument type (arguments in favor or against an issues) and the misclassification percentages, taking into consideration the best performing model for this task (Gemini1.5-flash, 5-shot). The complete per-argument evaluation results are reported in the Appendix C (Tables 9 to 14).

Specific, well-defined arguments are easier to detect Across both data sets, more specific arguments which focus on rights or concrete statements yield higher results. Examples include "*Removing under God would promote religious tolerance* (UG)" (F1: 0.74), "*Separation of state and religion* (UG)" (F1: 0.78) and "*Gay couples can declare their union without resort to marriage* " (F1: 0.70). More vague or abstract arguments, such as "*Implies ultimate power on the part of the state* (UG)" (F1: 0.66) and "*Marriage is about more than procreation* (GM)" (F1: 0.60) which allude to moral values rather than concrete statements were harder to predict for the models. We hypothesize that comments express those moral arguments in more diverse words, leading to a higher demand in abstraction to link the text back to the argument.

Arguments related to established institutions and legality We find a noticeable tendency for the models to perform better on arguments that reference established concepts or institutions. For example, arguments involving religion ("*Major world religions are against gay marriages* (GM)", F1: 0.75) and tradition ("'*Under God' is part of American tradition and history* (UG)", F1: 0.78) show relatively high F1 scores. Furthermore, arguments involving legal or rights-based concepts show moderate to high performance (e.g., "*Gay people should have the same rights as straight people* (GR; F1: 0.67)", *"It is discriminatory to refuse gay couples the right to marry* (GM)", *"Prohibition violates human rights* (MA; F1: 0.73)", "*Legalized marijuana can be controlled and regulated by the government* (MA; F1: 0.77)".

LLMs tend to overpredict arguments. In our exploratory analysis of LLMs, we observed a clear tendency for these models to overpredict positive instances when compared to the RoBERTa baseline across all three LLMs and across 0-, 1- and 5-shot settings. This is evidenced by the higher rates of false positive predictions (see Figure 3 to 6

in Appendix C). For example, the best-performing model, Gemini 1.5-flash (in the 5-shot setting), reported a false positive rate of 22%, significantly higher than the RoBERTa model, which only exhibited 2% false positives. Conversely, RoBERTa exhibits the opposite trhends of underpredition, namely a higher number of false negatives (77%) compared to 24% for Gemini. While all models reported higher number of false positives than the baseline, LLama had the highest percentage (42%), while the best-performing GPT model (1-shot) on this task reported a much higher percentage of false negatives than the other two LLMs (although still lower than RoBERTa). Beyond more specific instruction to combat the overprediction tendency in LLMs, ensemble models could be considered in future work given the complementary biases of LLMs and RoBERTa.

5 Conclusion

We have presented a systematic investigation of how well LLMs can identify the presence of arguments in a text, the identification of the exact span in which is it present, and whether the comment supports or attacks the argument. We used a controlled testbed of texts associated with predefined arguments that are commonly used in the discussion of the topics presented. Differently from typical AM tasks that deal with the identification of argumentative units and relationships, our approach allowed for a more-nuanced and systematic review of LLMs performance on topical arguments — a prerequisite for using them for large-scale argumentation analysis tasks.

First, we found that larger LLMs do not necessarily outperform traditional methods across all tasks. This was particularly true for Task 2 (Argument Extraction), where the best-performing LLM reached a comparable performance in the 5-shot setting with the fine-tuned RoBERTa model on the same task. All other models performed worse than the established baseline. Our results in this sense are in line with previous investigations on argument mining tasks with a prompting approach, in which LLMs did not surpass the established baselines (Al Zubaer et al., 2023; Ruiz-Dolz and Lawrence, 2023). While all LLMs outperformed the fine-tuned RoBERTa baselines in identifying the presence of an argument and the binary relationship between comment and argument (Task 1 and 3 respectively), performance dropped across all models when considering the full relationship scale in Task 3 (similar to the original work of Boltužić and Šnajder (2014)).

Contrary to our expectations, we found that the argument identification task was on average more challenging for the models than the binary relationship identification, meaning the models seem to understand better *how* a comment uses an argument, than *whether* the argument is used at all in the comment. We also found that introducing examples through one-shot or few-shot learning techniques did not consistently improve the performance of LLMs on the tasks.

Finally, our exploratory analysis highlighted that well-defined arguments are easier for the models to classify than vague, poorly-defined arguments. Our chosen LLMs here also performed consistently well on average on arguments related to established institutions, such as religion and legal rights. Previous research indicated that SOTA LLMs tend to favor liberal viewpoints (e.g., granting gay people the right to marry) over conservative ones (e.g., marriage should be between a man and a woman) (Tlaie, 2024; Pit et al., 2024). Our exploratory analysis revealed some degree of skew towards rights-centered arguments, and thus call for further investigation. We also found that, in comparison with more traditional methods (RoBERTa), LLMs tend to overpredict the presence of arguments in text, which raises questions about the reliability of LLM-based approaches in AM-related tasks.

Overall, our results call for caution when using LLMs to study opinions and ideologies with an argument-based approach. We show that although LLMs have been shown to perform well on traditional AM tasks, such as relation-based AM (Gorur et al., 2024; Otiefy and Alhamzeh, 2024), they exhibit systematic biases in recognition of different classes of arguments, which can lead to harmful misrepresentations in the model output. Finally, LLMs do not consistently benefit from added examples, and tend to overpredict. Solely relying on LLM prompting techniques to identify and scale the study argumentation might therefore lead to inaccurate classification. Future work is needed to understand the extent to which weaknesses can be addressed through better prompting; and perform follow-up analysis on the causes and effects of disparate accuracies on different classes of arguments.

6 Limitation

The data used in this study is limited in scope, both in terms of size and the range of topics and arguments it covers. While this controlled data set enabled a detailed analysis of Large Language Models (LLMs) in topical argumentation tasks, it may not fully represent the complexity and diversity of real-world argumentative discourse. Notably, the data sets employed were released in 2014, and may not capture more recent arguments or shifts in public opinion. For instance, the arguments related to the subtopic of gay marriage might no longer be relevant, especially given the legalization of gay marriage in the US in 2015, shortly after the data was released. On account of the limited data set size, we needed to conflate all datapoints for Task 1 to fine-tune our RoBERTa baseline, which made it impossible to compare the fine-tuned model with LLMs on each subtopic. Furthermore, due to cost and time constraints, we were not able to fine-tune any of the LLMs on the tasks. For these reasons, we also only employed GPT4o-mini, a much more compact model of the GPT4 family and than the other two models analyzed here.

7 Ethical Considerations

This study investigates the performance of LLMs in AM-related tasks on polarizing topics, which may involve sensitive or controversial discussions. We emphasize that the findings and conclusions of this research are not intended to amplify or legitimize harmful, discriminatory, or unethical viewpoints. Instead, the goal is to evaluate and enhance the understanding of LLMs' capabilities in argument detection, classification and extraction. Our research does not seek to endorse divisive or harmful opinions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*, 6:1278796.

Sarah M. Alsubhi, Areej M. Alhothali, and Amal A. Al-Mansour. 2023. AraBig5: The Big Five Personality Traits Prediction Using Machine Learning Algorithm on Arabic Tweets. *IEEE Access*, 11:112526–112534.

S. S. Arumugam. 2022. Development of argument based opinion mining model with sentimental data analysis from twitter content. *Concurrency and Computation: Practice and Experience*, 34(15):e6956. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.6956.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia. Association for Computational Linguistics.

T. J. M. Bench-Capon. 2003. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):429–448.

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.

Sarah Brown, Peter Anderson, and David Miller. 2023. Understanding the role of sampling parameters in language model generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3456–3470.

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. 2024. Exploring the Potential of Large Language Models in Computational Argumentation. *arXiv preprint*. ArXiv:2311.09022 [cs].

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.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Adrian de Wynter and Tangming Yuan. 2024. "I'd Like to Have an Argument, Please": Argumentative Reasoning in Large Language Models. In *Computational Models of Argument*, pages 73–84. IOS Press.

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

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2020. Corpus for Modeling User Interactions in Online Persuasive Discussions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1135–1141, Marseille, France. European Language Resources Association.

Tarek Elghazaly, Amal Mahmoud, and Hesham A Hefny. 2016. Political sentiment analysis using twitter data. In *Proceedings of the International Conference on Internet of things and Cloud Computing*, pages 1–5.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2024. Can Large Language Models perform Relation-based Argument Mining? *arXiv preprint*. ArXiv:2402.11243 [cs].

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Martin Hinton and Jean HM 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 & Computation*, 14(1):59–74.

Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. 2010. Mining economic sentiment using argumentation structures. In *Advances in Conceptual Modeling–Applications and Challenges: ER 2010 Workshops ACM-L, CMLSA, CMS, DE@ ER, FP-UML, SeCoGIS, WISM, Vancouver, BC, Canada, November 1-4, 2010. Proceedings 29*, pages 200–209. Springer.

Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. 2023. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290.

John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO. Association for Computational Linguistics.

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

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

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Yanting Liu, Xue Zhang, and Brian Thompson. 2023. An empirical study of temperature parameter impact on large language model outputs. *Transactions of the Association for Computational Linguistics*, 11:845–862.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaolong Meng, Jianxin Wu, and Kai Chen. 2023. Enhancing reproducibility in large language models: A study of temperature and top-p parameters. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1123–1135.

Rafael Mestre, Matt Ryan, Stuart E Middleton, Richard Gomer, Masood Gheasi, Jiatong Zhu, and Timothy J Norman. 2022. Benchmark evaluation for tasks with highly subjective crowdsourced annotations: Case study in argument mining of political debates.

Yasser Otiefy and Alaa Alhamzeh. 2024. Exploring Large Language Models in Financial Argument Relation Identification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 119–129, Torino, Italia. Association for Computational Linguistics.

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

Pagnarasmey Pit, Xingjun Ma, Mike Conway, Qingyu Chen, James Bailey, Henry Pit, Putrasmey Keo, Watey Diep, and Yu-Gang Jiang. 2024. Whose side are you on? investigating the political stance of large language models. *ArXiv*, abs/2403.13840.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most. Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh. 2016. Election result prediction using twitter sentiment analysis. In *2016 international conference on inventive computation technologies (ICICT)*, volume 1, pages 1–5. IEEE.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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

Ramon Ruiz-Dolz, John Lawrence, Ella Schad, and Chris Reed. 2024. Overview of DialAM-2024: Argument Mining in Natural Language Dialogues. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 83–92, Bangkok, Thailand. Association for Computational Linguistics.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From Argumentation Mining to Stance Classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO. Association for Computational Linguistics.

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

Manfred Stede and Jodi Schneider. 2018. *Argumentation mining*. Springer.

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25.

Alejandro Tlaie. 2024. Exploring and steering the moral compass of large language models. *arXiv preprint arXiv:2405.17345*.

Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *HHAI2022: Augmenting Human Intellect*, pages 17–31. IOS Press.

Mei Zhang, Wei Chen, Yixuan Wang, and Hongzhi Li. 2023. Investigating the impact of decoding strategies on large language model performance: A systematic analysis. *arXiv preprint arXiv:2306.09265*.

A Lists of Arguments

Here, we present the complete list of pro and con arguments from the original datasets in Table 4.

B Prompts

We display the prompts used for our three tasks in Table 5 to Table 8.

C Detailed Results

Additionally, Table 9 to Table 14 report the full metrics for each subtopic for the per-argument analysis for the best-performing model in Task 1, as explained in Section 4.4. Finally, Figure 3 to Figure 6 contain the confusion matrices in percentage from Task 1 for the combined-data fine-tuned RoBERTa and for GPT, Gemini and Llama (the classification reported for each model corresponds to the best performing shot-setting).

Data set	Pro Arguments	Con Arguments
GM	It is discriminatory to refuse gay couples the right to marry. Gay couples should be able to take advantage of the fiscal and legal benefits of marriage. Marriage is about more than procreation, therefore gay couples should not be denied the right to marry due to their biology. Others	Gay couples can declare their union without resort to marriage. Gay marriage undermines the institution of marriage, leading to an increase in out-of-wedlock births and divorce rates. Major world religions are against gay marriages. Marriage should be between a man and a woman. Others
UG	Likely to be seen as a state-sanctioned condemnation of religion. The principles of democracy regulate that the wishes of American Christians, who are a majority, are honored. "Under God" is part of the American tradition and history. Others	Implies ultimate power on the part of the state. Removing "under God" would promote religious tolerance. Separation of state and religion. Others
AB	Abortion is a woman's right. Rape victims need it to be legal. A fetus is not a human yet, so it's okay to abort. Abortion should be allowed when a mother's life is in danger. Unwanted babies are ill-treated by parents and/or not always adopted. Birth control fails at times, and abortion is one way to deal with it. Abortion is not murder. Mother is not healthy/financially solvent. Others	Put the baby up for adoption. Abortion kills a life. An unborn baby is a human and has the right to live. Be willing to have the baby if you have sex. Abortion is harmful to women. Others
GR	Gay marriage is like any other marriage. Gay people should have the same rights as straight people. Gay parents can adopt and ensure a happy life for a baby. People are born gay. Religion should not be used against gay rights. Others	Religion does not permit gay marriages. Gay marriages are not normal/against nature. Gay parents cannot raise kids properly. Gay people have problems and create social issues. Others
MA	Not addictive. Used as a medicine for its positive effects. Legalized marijuana can be controlled and regulated by the government. Prohibition violates human rights. Does not cause any damage to our bodies. Others	Damages our bodies. Responsible for brain damage. If legalized, people will use marijuana and other drugs more. Causes crime. Highly addictive. Others
OB	Fixed the economy. Ending the wars. Better than the Republican candidates. Makes good decisions/policies. Has qualities of a good leader. Ensured better healthcare. Executed effective foreign policies. Created more jobs. Others	Destroyed our economy. Wars are still ongoing. Unemployment rate is high. Healthcare bill is a failure. Poor decision-maker. We have better Republicans than Obama. Not eligible as a leader. Others

Table 4: Pro and Con Arguments for All Subtopics and Data Sets

Analyze the given comment about `topic` in relation to a specific argument. You need to:
Identify if the comment makes use of the given argument. If it does, assign the label 1. If it does not, assign the label 0. Do NOT use any other label.
The argument to analyze is: `argument`.
Provide your response in the following JSON format:

```
"id": "id", "label": "the label for the use of the argument in the comment"
```

Analyze the following comment in relation to the given argument: comment

Table 5: Prompt for Task 1

Analyze the given comment about `topic` in relation to a specific argument. You need to:
Identify if the comment makes use of the given argument. Assign the following labels:
1 if the comment attacks the argument. 5 if the comment supports the argument. Do NOT use any other label. Do NOT include the comment or the argument in the response. The argument to analyze is: `argument`.
Provide your response in the following JSON format:

```
"id": "id", "label": "the label for the use of the argument in the comment"
```

Analyze the following comment in relation to the given argument: comment

Table 6: Prompt for Task 2 - binary

Analyze the given comment about `topic` in relation to a specific argument. You need to:
Identify if the comment makes use of the given argument. Assign the following labels:
1 if the comment attacks the argument explicitly. 2 if the comment attacks the argument implicitly/vaguely. 4 if the comment supports the argument implicitly/vaguely. 5 if the comment supports the argument explicitly. Do NOT use any other label. Do NOT include the comment or the argument in the response.
The argument to analyze is: `argument`.
Provide your response in the following JSON format:

```
"id": "id", "label": "the label for the use of the argument in the comment"
```

Analyze the following comment in relation to the given argument: comment

Table 7: Prompt for Task 2 - full scale

Analyze the given comment about `topic` in relation to a specific argument. You need to:
Identify the relevant span of text where the comment makes use of the given argument. Provide the exact span of the text in the comment that makes use of the argument. Only report the exact original fragment of text. Do NOT paraphrase. Do NOT include any additional text.
The argument to analyze is: `argument`.
Provide your response in the following JSON format:

```
"id": "id", "span": "the relevant span of text"
```

Analyze the following comment in relation to the given argument: comment

Table 8: Prompt for Task 3

Argument	P	R	F1
Gay couples can declare their union without resort to marriage	0.62	0.81	0.70
Gay couples should be able to take advantage of the fiscal and legal benefits of marriage	0.60	0.71	0.65
Gay marriage undermines the institution of marriage, leading to an increase in out-of-wedlock births and divorce rates	0.54	0.61	0.57
It is discriminatory to refuse gay couples the right to marry	0.74	0.64	0.68
Major world religions are against gay marriages	0.79	0.71	0.75
Marriage is about more than procreation, therefore gay couples should not be denied the right to marry due to their biology	0.58	0.62	0.60
Marriage should be between a man and a woman	0.56	0.55	0.55

Table 9: Gemini 5-shot results for GM, Task 1

Argument	P	R	F1
Separation of state and religion	0.78	0.54	0.64
Removing "Under God" would promote religious tolerance	0.71	0.77	0.74
"Under God" is part of American tradition and history	0.79	0.77	0.78
Likely to be seen as a state-sanctioned condemnation of religion	0.49	0.47	0.48
Implies ultimate power on the part of the state	0.52	0.91	0.66
America is based on democracy and the pledge should reflect the belief of the American majority	0.57	0.54	0.55

Table 10: Gemini 5-shot results for UGIP, Task 1

Argument	P	R	F1
A fetus is not a human yet, so it's okay to abort	0.72	0.76	0.74
Abortion should be allowed when a mother's life is in danger	0.89	0.91	0.90
Unwanted babies are ill-treated by parents and/or not always adopted	0.76	0.78	0.77
Birth control fails at times and abortion is one way to deal with it	0.65	0.81	0.72
Abortion is not murder	0.55	0.69	0.61
Mother is not healthy/financially solvent	0.69	0.76	0.72
Put baby up for adoption	0.77	0.91	0.84
Abortion is a woman's right	0.73	0.79	0.76
Rape victims need it to be legal	0.83	0.95	0.89
Abortion kills a life	0.64	0.65	0.65
An unborn baby is a human and has the right to live	0.63	0.64	0.64
Be willing to have the baby if you have sex	0.75	0.87	0.80
Abortion is harmful for women	0.54	0.86	0.66
Others (pro)	0.47	0.49	0.48
Others (con)	0.57	0.53	0.55

Table 11: Gemini 5-shot results for Abortion, Task 1

Argument	P	R	F1
Gay people should have the same rights as straight people	0.69	0.66	0.67
Gay parents can adopt and ensure a happy life for a baby	0.75	0.84	0.79
People are born gay	0.76	0.66	0.71
Religion should not be used against gay rights	0.58	0.71	0.64
Gay marriage is like any other marriage	0.58	0.64	0.61
Religion does not permit gay marriages	0.64	0.83	0.72
Gay marriages are not normal/against nature	0.67	0.81	0.73
Gay parents can not raise kids properly	0.61	0.82	0.70
Gay people have problems and create social issues	0.63	0.80	0.71
Others (pro)	0.44	0.49	0.46
Others (con)	0.52	0.51	0.51

Table 12: Gemini 5-shot results for Gay Rights, Task 1

Argument	P	R	F1
Not addictive	0.68	0.69	0.68
Used as a medicine for its positive effects	0.67	0.77	0.72
Legalized marijuana can be controlled and regulated by the government	0.76	0.78	0.77
Prohibition violates human rights	0.69	0.77	0.73
Does not cause any damage to our bodies	0.62	0.70	0.65
Damages our bodies	0.62	0.82	0.70
Responsible for brain damage	0.62	0.82	0.70
If legalized, people will use marijuana and other drugs more	0.55	0.68	0.61
Causes crime	0.56	0.77	0.65
Highly addictive	0.63	0.73	0.68
Others (pro)	0.53	0.55	0.54
Others (con)	0.53	0.55	0.54

Table 13: Gemini 5-shot results for Marijuana, Task 1

Argument	**P**	**R**	**F1**
Fixed the economy	0.75	0.67	0.71
Ending the wars	0.65	0.77	0.71
Better than the republican candidates	0.75	0.8	0.77
Makes good decisions/policies	0.57	0.69	0.63
Has qualities of a good leader	0.61	0.77	0.68
Ensured better healthcare	0.82	0.79	0.8
Executed effective foreign policies	0.62	0.72	0.67
Created more jobs	0.6	0.68	0.64
Destroyed our economy	0.68	0.81	0.74
Wars are still on	0.64	0.8	0.71
Unemployment rate is high	0.62	0.79	0.69
Healthcare bill is a failure	0.58	0.79	0.67
Poor decision-maker	0.54	0.69	0.61
We have better republicans than Obama	0.6	0.63	0.61
Not eligible as a leader	0.64	0.79	0.71
Ineffective foreign policies	0.53	0.76	0.62
Others (pro)	0.46	0.5	0.48
Others (con)	0.45	0.49	0.47

Table 14: Gemini 5-shot results for Obama, Task 1

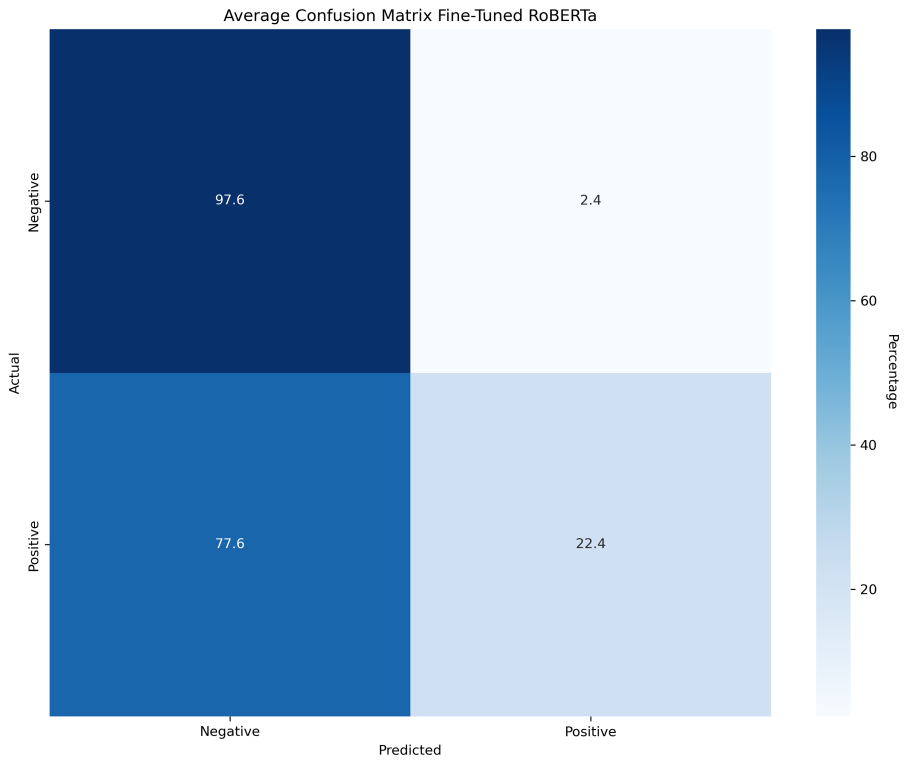

Figure 3: Misclassification report (percentages) for the fine-tuned RoBERTA, Task 1

.

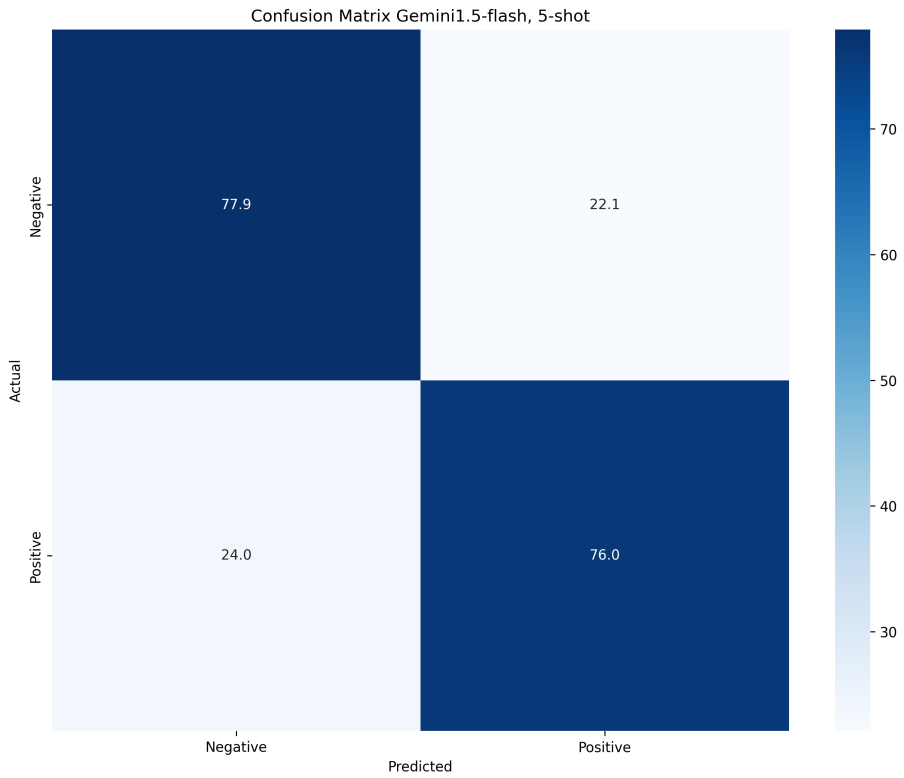

Figure 4: Misclassification report (percentages) for Gemini1.5-flash, 5-shot, Task 1

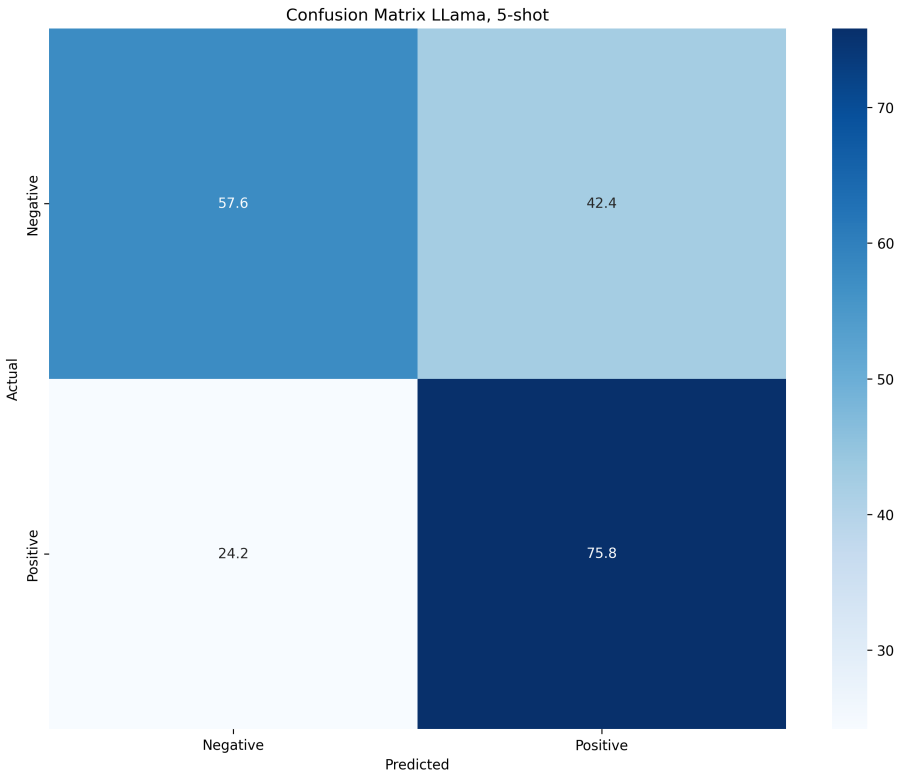

Figure 5: Misclassification report (percentages) for Llama3-8b, 5-shot, Task 1

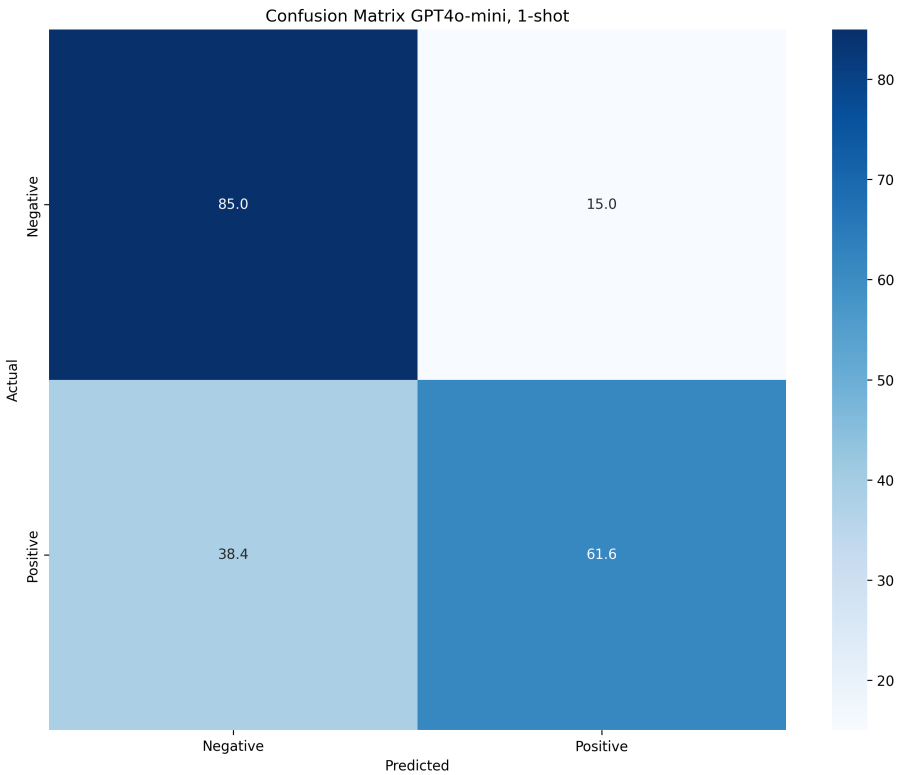

Figure 6: Misclassification report (percentages) for GPT4o-mini, 1-shot, Task 1