Assessing Open-Weight Large Language Models on Argumentation Mining Subtasks

Mohammad Yeghaneh Abkenar[*1,3]
yeghanehabkenar@uni-potsdam.de

Weixing Wang[*2,3]
weixing.wang@hpi.de

Hendrik Graupner[1,2,3]
hendrik.graupner@hpi.de

Manfred Stede[3]
stede@uni-potsdam.de

[1]Bundesdruckerei Gruppe GmbH Berlin, [2]Hasso Plattner Institute, [3]University of Potsdam

Abstract

We explore the capability of four open-weight large language models (LLMs) in argumentation mining (AM). We conduct experiments on three different corpora; persuasive essays (PE), argumentative microtexts (AMT) Part 1 and Part 2, based on two argumentation mining subtasks: (i) argument component type classification (ACTC), and (ii) argumentative relation classification (ARC). This work aims to assess the argumentation capability of open-weight LLMs, including Mistral 7B, Mixtral 8x7B, LLaMA2 7B and LLaMA3 8B in both, zero-shot and few-shot scenarios. Our results demonstrate that open-weight LLMs can effectively tackle argumentation mining subtasks, with context-aware prompting improving relation classification performance, though the models' effectiveness varies across different argumentation patterns and corpus types, suggesting potential for specialized adaptation in future argumentation systems. Our analysis advances the assessment of computational argumentation capabilities in open-weight LLMs and provides a foundation for future research.[1]

1 Introduction

Over the past few years, advancements in the broader field of natural language processing (NLP), such as pre-trained transformer-based models (Devlin, 2018), coupled with the increasing availability of diverse data, have significantly enhanced the potential for nearly every area of NLP, including argumentation mining (AM) (Stede and Schneider, 2018; Lawrence and Reed, 2020). AM, and specifically the problem of finding argumentation structures in text, has received much attention in the past decade. The objective of AM is to detect argumentation within text or dialogue, to create detailed representations of claims and their supporting or attacking arguments, and to analyze the reasoning patterns that validate the argumentation. Beyond academic interest, AM attracts significant attention for its diverse applications, as demonstrated by projects like IBM Debater (Bar-Haim et al., 2021), decision assistance (Liebeck et al., 2016), product reviews (Passon et al., 2018) and writing support (Wachsmuth et al., 2016).

[1]Code and data available on https://github.com/myeghaneh/OpenArgMinLLM/tree/main

2 Background and Related work

2.1 Argumentation Mining

Unlike many NLP problems, argumentation mining (AM) is not a single, straightforward task but rather a collection of interrelated subtasks. AM enhances sentiment analysis by delving deeper into the reasoning behind opinions. While sentiment analysis identifies "*what* people think about entity X," AM explores "*why* people think Y about X." One subtask we address is argument component type classification (ACTC), which identifies the type of argumentative discourse units, as defined by Hidey et al. (2017, p. 14) as follows:

- *Claim* (Conclusion): A statement in the text that articulates a perspective on a particular issue. It can include predictions, interpretations, evaluations, and expressions of agreement or disagreement with others' assertions.

- *Premise* (Evidence): A statement presented to strengthen a claim, designed to persuade the audience of its validity. Although premises may express opinions, their main function is to support or refute an existing proposition rather than introduce a new perspective.

We also cover argumentative relation classification (ARC) to identify relations among argumentative discourse units (ADUs) which is defined by Ali et al. (2022, p. 491) as follows:

- *Support* (For): The Support relation occurs when a premise enhances or reinforces a claim.

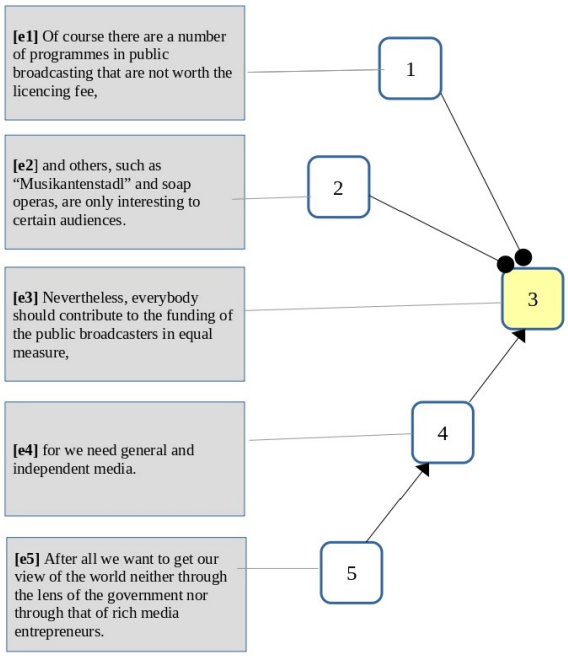


Figure 1: A simplified example from (Peldszus and Stede, 2015b), argumentative microtexts corpus, part 1. The argument structure consists of five elements (e1–e5) with argumentation component type annotation; premise (white boxes) and claim (yellow box) nodes, and supporting (arrow-head) and attacking (circle-head) relations

This can happen in various ways: If the claim is a proposition (such as a fact, opinion, or belief), the premise strengthens the claim's likelihood or truth. If the claim is an action, the premise provides justification or makes the action more acceptable. If the claim is an event, the premise increases the probability that the event occurred.

- *Attack* (Against): The Attack relation occurs when a premise undermines or contradicts a claim. This can manifest in several ways: If the claim is a proposition, the premise weakens the claim's likelihood or truth. If the claim is an action, the premise denies or challenges the justification for the action. If the claim is an event, the premise reduces the probability that the event occurred.

  Figure 1 provides a simplified example from our corpus, showcasing their component types and relationships. It also focuses on two subtasks and their interconnection, demonstrating how they are related and work together to form a final argument.

2.2 Using LLMs for AM

Recently, we saw huge breakthroughs in language modeling. Large Language Models such as GPT4 (Achiam et al., 2023), LLaMA3 (Dubey et al., 2024b), and Mistral (Jiang et al., 2023) have demonstrated strong capabilities in solving various NLP tasks. LLMs are capable of capturing the nuances, context, and semantics of the human language, allowing them to perform tasks such as text generation (Zhao et al., 2023), summarization (Jin et al., 2024; Chang et al., 2023; Zhang et al., 2024), translation (Wu et al., 2024; Xu et al., 2024; Li et al., 2024a), question answering (Li et al., 2024b; Wei et al., 2022), and more. As a result, there is an increasing interest in applying LLMs for computational argumentation tasks. For example, de Wynter and Yuan (2023) evaluated the ability of two LLMs to perform argumentative reasoning. Their experiments involved argumentation mining and argument pair extraction, assessing the LLMs' capability to recognize arguments under progressively more abstract input and output representations. However, their research is limited to the two closed-source language models GPT3 and GPT4. Chen et al. (2023) conducted a comprehensive analysis of LLMs on diverse computational argumentation tasks, their goal was to evaluate LLMs including ChatGPT, Flan models, and LLaMA2 models in both zero-shot and few-shot settings. However, their studies did not address the argumentative relation classification subtask and

they did not use some state-of-the-art models such as LLaMA3 and the Mistral family which according to Sinha et al. (2024) are also promising in various reasoning tasks.

To overcome the above limitations, we explore two key subtasks of argumentation: argumentation discourse unit classification and argument relation classification, using four open-source LLMs across three well-known argument mining corpora. We believe that argumentation mining subtasks are fundamentally different from argument pair extraction and argument generation. As such, argumentation mining subtasks need to be explored differently using various LLMs on the most well-known and important corpora with a similar structure.

3 Corpora and Task Definition

One approach to assessing the reasoning capabilities of LLMs is to evaluate concretely their performance on tasks that necessitate reasoning. We have chosen this approach, in order to measure the ability of different large language models in reasoning. In this paper, we conduct experiments on two central subtasks of argumentation mining using three well-known datasets ,which will be introduced in the next subsections.

3.1 Corpora

Dataset/Subtask	ACTC			ARC		
	Total	Claim	Premise	Total	Support	Attack
AMT1	576	112	464	455	284	171
AMT2	932	171	761	738	524	214
PE	6089	2257	3832	3821	3603	218

Table 1: Summary of sample number and label distributions of the three corpora.

Argumentative Microtexts Part 1(AMT1) The AMT1 corpus, created by (Peldszus and Stede, 2015a), includes 112 short texts (each about 3–5 sentences long) and 576 argumentative discourse units. They were originally written in German and have been professionally translated to English, as well as to Italian (Namor and Stede, 2019), Russian (Fishcheva and Kotelnikov, 2019) and recently to Persian (Abkenar and Stede, 2024) preserving the segmentation and if possible the usage of discourse markers and annotated with complete argumentation tree structures.

Argumentative Microtexts Part 2(AMT2). The second part of AMT, created by (Skeppstedt et al., 2018) using crowd-sourcing, includes 171 short texts with 932 argumentative discourse units in English which is annotated consistent with the approach utilized in the original corpus. One of the differences in this corpus is the existence of an implicit claim which is marked in the XML file.

Persuasive Essays(PE) The PE corpus comprises 402 argumentative essays (totaling 2235 paragraphs) written by English learners in response to specific prompts. Stab and Gurevych (2017) collected these essays from a website and annotated them with argumentation graphs. The essays begin with a question and include a major claim supported by evidence, which may have a substructure. Some sentences are non-argumentative, providing only background or minor elaborations. Each essay has a major claim, typically found at the end, supported by claims within the paragraphs. For consistency with other corpora, we treat "major claim" and "claim" as equivalent and classify argument components (ACs) at the paragraph level.

3.2 Tasks

Argument Component Type Classification (ACTC) Argumentative discourse units (ADUs) are minimal units of analysis, i.e., the smallest elements in a text that contribute to argumentative structure. In this paper, we define ACTC as the classification of these units as either "premise" or "claim"; we do not address the distinction between ADUs and non-argumentative material.

Argumentative Relation Classification(ARC) The goal of argumentative relation identification is to determine whether each pair of ADUs is argumentatively related or not (Rocha et al., 2018). We assume that the task of segmenting the text into ADUs has already been completed. Following (Stab and Gurevych, 2014), given an ordered pair of ADUs, the objective is to classify the relation between them as either "support" or "attack."

4 Methods

4.1 Vanilla Prompting

This approach involves asking the model to classify each ADU independently, without considering the whole context. As shown on the left side of Figure 2, we ask the model: "Please classify the following ADU q_i into one of the categories C_i." This is the same for the ARC, but we ask the same question on pairs of ADUs.

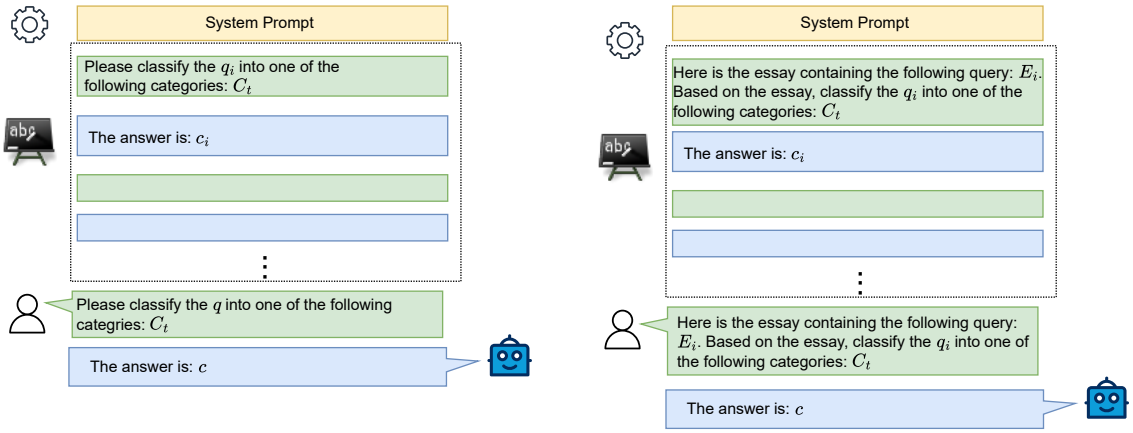


Figure 2: An overview of the prompting methods. Left: Vanilla Prompting. Right: Context-Aware Prompting

4.2 Context-Aware Prompting

This approach asks the model to classify each ADU based on its context in the text. As shown on the right side of Figure 2, we prompt the model with:

"Here is the essay containing the following query E_i. Based on the essay, classify q_i into one of the categories C_i." Unlike the standard method, where each ADU is classified independently, this context-aware prompting requires the model to consider the surrounding context of the essay or microtext for each ADU. For ARC, we ask the model to classify pairs of ADUs, still taking into account the context provided.

4.3 Prompt Design and Engineering

We designed our prompts to clearly communicate the task requirements while avoiding unnecessary complexity. For both vanilla and context-aware approaches, we provided the model with a system message identifying it as "an expert in linguistics and argumentation mining" to prime it for the specialized task.

For few-shot learning, we carefully selected demonstration examples to represent balanced class distributions and varying difficulty levels. When constructing demonstrations, we ensured that they represented diverse argumentative patterns and linguistic constructions present in the target corpus.

We performed preliminary experimentation to optimize prompt formatting, including the use of explicit indicators like "The answer is:" to guide the model's output format. This standardization facilitated easier evaluation and reduced parsing errors. Please refer to appendix B for the design of full prompts for all tasks.

4.4 Model Selection and Implementation

We test our two prompting methods with 4 advanced LLMs, namely **LLaMA 2-7B** (Touvron et al., 2023), **LLaMA 3-8B** (Dubey et al., 2024a), **Mistral-7B** (Jiang et al., 2023), and **Mixtral-8x7B** (Jiang et al., 2024). All samples for all three corpora are within the context window of each model. For comparison, we report the micro F_1-score, because the datasets are all imbalanced.

For implementation, we used the Hugging Face Transformers library[2] to access these models, running inference with a batch size of 1 and random seeds to minimize randomness in outputs. All experiments were conducted three times on NVIDIA A100 GPUs with 40GB of memory. We implemented automated post-processing of model outputs to extract predicted labels and compute metrics for evaluation. As shown in the highlighted part of the figures 3 and 4, performance remained consistently stable, showing minimal variation from stochastic effects.

5 Experiments and Results

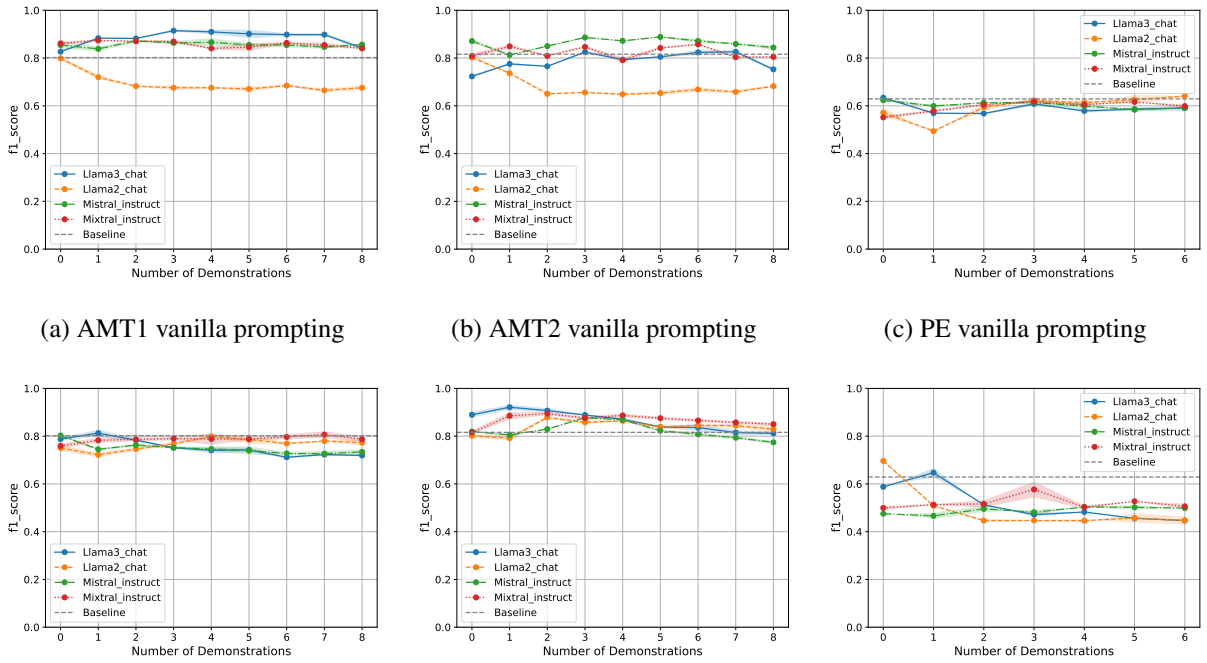
5.1 Baseline

For the experiments on ACTC, we employ a simple strategy of predicting the most frequent (majority) type observed for each ADU type in each of the corpora. As seen in the last row of the table 2 , this approach results in micro F_1-scores of 0.802 for AMT1, 0.816 for AMT2 and 0.629 for PE. Moreover, for the experiment on ARC, we followed the same strategy to calculate a baseline on relation types. This gave us: a micro F_1-score of 0.624 for MT1, 0.710 for MT2 and 0.942 for PE.

[2] https://huggingface.co/docs/transformers/de/

Model	AMT1				AMT2				PE			
	ACTC		ARC		ACTC		ARC		ACTC		ARC	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Mistral (Vanilla)	0.854	**0.745**$^+$	0.491	0.219	0.656	0.767	0.578	0.367	0.623	0.510	0.724	0.340
Mistral (Context)	0.802	0.463	**0.651**$^+$	**0.262**	0.456	0.502	0.693	0.434	0.475	0.428	0.792	0.370
Mixtral (Vanilla)	**0.861**$^+$	0.728	0.556	0.238	0.566	0.604	0.638	0.390	0.551	0.543	0.784	0.371
Mixtral (Context)	0.759	0.585	0.604	0.251	0.598	0.674	**0.734**$^+$	**0.451**$^+$	0.499	0.543	0.887	0.439
LLaMA2 (Vanilla)	0.798	0.489	0.216	0.118	0.465	0.471	0.291	0.236	0.571	0.544	0.350	0.210
LLaMA2 (Context)	0.750	0.574	0.017	0.002	0.577	0.656	0.222	0.154	**0.696**$^+$	0.546	0.632	0.270
LLaMA3 (Vanilla)	0.826	0.717	0.222	0.181	0.514	0.518	0.703	0.380	0.634	**0.612**$^+$	0.883	**0.583**$^+$
LLaMA3 (Context)	0.787	0.657	0.302	0.213	**0.671**	**0.816**$^+$	0.719	0.387	0.588	0.469	**0.931**	0.428
Majority Baseline	0.802	0.446	0.624	0.384	0.816	0.449	0.710	0.415	0.629	0.386	0.942	0.485

Table 2: Performance of different models across AMT1, AMT2 and PE corpora on ACTC, and ARC tasks. The bold values in the table represent the best result for each subtask and dataset, while the $^+$ indicates which of these results were able to outperform the baseline in zero-shot settings



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting

(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

Figure 3: ACTC. The first row shows the model performance using vanilla prompting on three datasets whereas the below row shows the performance with the context-aware prompting.

5.2 Zero-Shot Performance

Table 2 presents the results of zero-shot prompting. When comparing ACTC and ARC tasks, we find that context prompts generally improve ARC performance across most models and datasets, suggesting that context aids in better understanding the relationships between sentences. However, the ACTC task appears more sensitive to the introduction of context, with some models experiencing a performance drop. LLaMA3 stands out for maintaining strong performance across both tasks and all datasets when using context prompts. This suggests that LLaMA3 is more adaptable to varying prompting methods and datasets in AM tasks.

Considering these baselines, it is clear that the results on ARC are significantly better than those on ACTC, which could be due to the differences in task definitions and their subjective nature. For instance, identifying a claim within a text may be more subjective and context-dependent, requiring a deeper understanding of the argument. In contrast, determining whether two ADUs are supporting or attacking each other is relatively more straightforward and less ambiguous, making it easier for LLMs to classify them.

Regarding evaluation metrics, we observe that LLMs sometimes underperform the baseline on Micro F1 scores while showing stronger results

on Macro F1. This pattern suggests dataset imbalance, where the majority class dominates the Micro metric calculations. The stronger Macro F1 performance indicates that LLMs are better at handling minority classes when evaluating across all classes equally.

Given that LLMs are known to be effective few-shot learners (Brown et al., 2020), these promising zero-shot results suggest significant potential for further optimization through few-shot learning approaches, which we explore in the following section.

5.3 Performance and Number of Demonstrations

5.3.1 Results on ACTC

In the following sections, we only focus on Micro F1 scores. The full results on Macro F1 can be found in A. Figure 3 illustrates the performance on the ACTC task across the datasets, under different numbers of demonstrations. For the ACTC task, context-aware prompting can bring all models to a similar level. I.e., weaker models like LLaMA2 are enhanced while the stronger models are degraded. For example, considering three-shot learning on the AMT1 dataset, LLaMA3 can achieve 86% micro F_1-score using vanilla prompting (a), but the micro F_1-score drops to 81% with context-aware prompting (d). For the AMT2 dataset, we observe similar phenomena in (b) and (e), however, here LLaMA3 achieved the best results in the first shot using context-aware prompting.

In comparison, LLaMA2 improves from 79% (a) to 87% (d) by applying context-aware prompting. Moreover, we find that the application of context-aware prompting significantly reduces the performance disparity between the AMT1 and AMT2 datasets. This suggests that providing additional contextual information helps the models to handle variations between these datasets more effectively, resulting in a more uniform performance across different versions of the AM tasks.

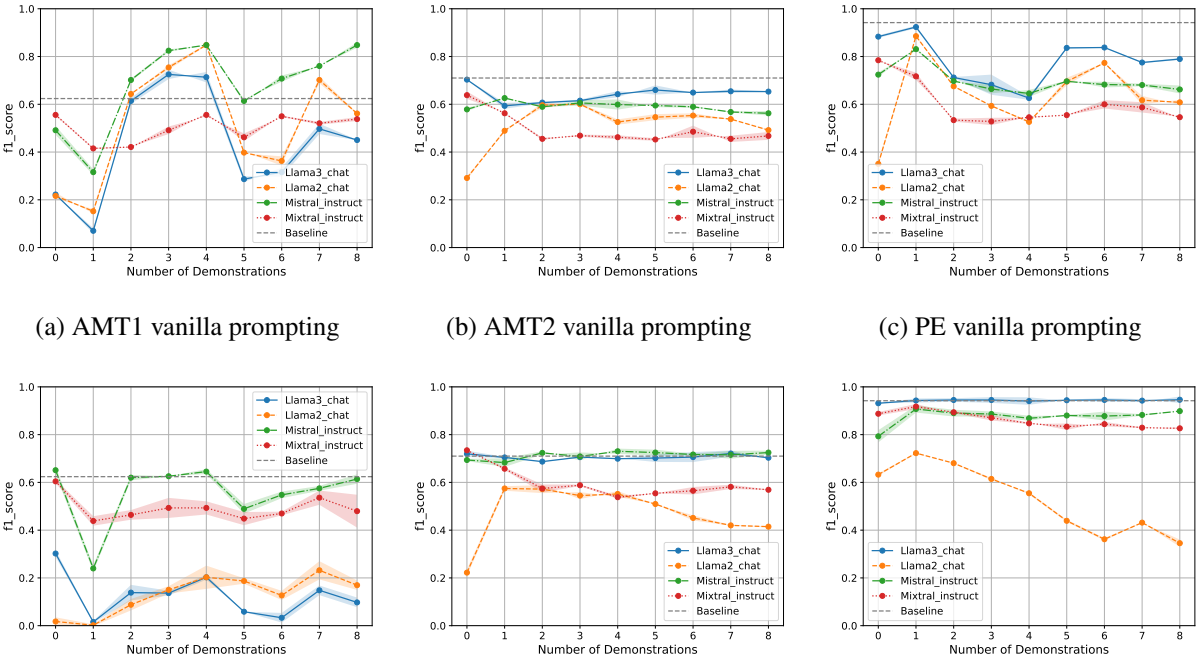
Our few-shot experiments on ACTC highlight the complexities of adapting to different argumentation styles. In the PE vanilla prompting setup in (c), model performance remains relatively stable across different numbers of demonstrations, with slight variations among models. This suggests that ACTC's argumentation structures may not be as easily influenced by increasing demonstration. However, in the context-aware prompting setting of PE, we see more fluctuations in (f), particularly in the early demonstrations. One possible explanation is the longer text length in PE dataset compared to the MT datasets. This corpus differs from the microtext corpora in that each paragraph can contain more than one claim, which impacts the weighting of component in the final F_1-micro score calculation. Furthermore, for the ACTC subtask in the PE dataset, the addition of contextual information could actually degrade the model's ability to solve the task effectively. The increased context could introduce more complexity, making it harder for models to solve ACTC task, which is one of subjective and complex task of Argumentation Mining that is align with finding in (Levy et al., 2024)

5.3.2 Results on ARC

For the ARC task, we see slightly different patterns of model performance in Figure 4. However, we still observe that context-aware prompting serves as an effective stabilizer for model performance. Comparing (a) and (d), we find that when models are prompted with additional contextual information, they exhibit reduced fluctuations in their performance regarding different numbers of demonstrations, suggesting that this approach helps mitigate the impact of noise brought by additional demonstrations. In contrast, vanilla prompting, which lacks this additional context, often results in more erratic performance across different numbers of demonstrations, likely because the models are more susceptible to the inherent variability and difficulty of the tasks. This fluctuation in vanilla prompting can be attributed to the models' struggle to consistently grasp the underlying patterns in the data without sufficient context, leading to inconsistent F_1-scores. By providing context-aware prompting, the models are better equipped to understand and process the tasks at hand, resulting in more stable and reliable outputs.

Our few-shot experiments with the ARC task highlight the challenge of transfer learning across different argumentation patterns. The PE corpus, with its academic writing style, showed the most consistent improvement with additional demonstrations, suggesting that formal argumentation patterns may be more learnable from examples. In contrast, the more varied AMT2 corpus showed less consistent improvement patterns, indicating that diverse argumentation styles may require more sophisticated adaptation approaches.

In comparing model architectures, we observed



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting

(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

Figure 4: ARC. The first row shows the model performance using vanilla prompting on three datasets where the below row shows the performance with the context-aware prompting.

that Mixtral's mixture-of-experts architecture consistently outperforms the others in the few-shot regime for relation classification, potentially due to its ability to activate different expert pathways for different relation types. This architectural advantage is particularly evident in the context-aware setting, where the model must integrate information across longer text spans.

5.4 Error Analysis and Qualitative Assessment

We conducted a detailed error analysis to understand when and why models fail at argumentation mining tasks. For ACTC, all models struggle most with claims that lack explicit stance indicators or that use hedging language. For instance, in AMT1, the sentence "Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins" was often misclassified as a claim due to its evaluative language, despite functioning as a premise in context.

For ARC, the most challenging cases involve implicit support or attack relations where no explicit discourse markers (like "because" or "however") are present. Models particularly struggle with relations that require domain knowledge to interpret correctly. Additionally, all models show a bias toward predicting the majority class, especially in the PE corpus, where support relations vastly outnumber attack relations.

Qualitatively, we observed that LLaMA3 produces more coherent explanations for its decisions when prompted to explain its reasoning, suggesting deeper understanding of argumentative structures. Mixtral exhibits greater sensitivity to subtle indicators of argumentative function, while Mistral performs better at identifying explicit discourse markers as indicators of relation type.

6 Discussion and Conclusion

6.1 Result Comparison with Literature

We assessed the reasoning abilities of four LLMs. Our evaluation focused on two sub-tasks in argumentation mining: ACTC and ARC. However, comparing these results with the state of the art is not straightforward, primarily due to the variations in how different metrics are evaluated and reported across studies. The LLMs performed particularly excelled in ARC in comparison to our majority baseline, and performed well in ACTC, surpassing or closely matching the results reported in (Abkenar et al., 2021) and (Chernodub et al., 2019) for AMT1 and PE, based on the micro F_1-score. However, statistical analysis of the LLMs' predictions shows that their performance generally

differs between AMT1 and AMT2, which we attribute to a difference in text quality due to the varying elicitation conditions. We also revealed that demonstrations serve as stabilizers rather than enhancers for both AM tasks.

6.2 Theoretical Implications

Our findings have several theoretical implications for understanding LLMs' capabilities in structured reasoning tasks. First, the models' strong zero-shot performance suggests they have acquired implicit knowledge of argumentation structures during pretraining, despite not being explicitly trained on argumentation tasks that we designed in this work. This supports the hypothesis that general language understanding includes some degree of argumentation comprehension. Second, the stabilizing rather than enhancing effect of demonstrations suggests that few-shot learning in AM primarily helps models understand task framing rather than teaching them new argument patterns. This challenges simplistic views of in-context learning as analogous to traditional learning from examples.

Limitations

We conducted our study on two central subtasks of AM. However, other subtasks, such as the identification of argument components and the evaluation of argument quality, need to be addressed. We also aim to evaluate more recent LLMs, such as DeepSeek (Guo et al., 2025) and Hermes (Teknium et al., 2024), which are potentially strong in reasoning. For future work, we intend to explore the impact of input length on model performance in AM subtasks. Additionally, our results focus exclusively on English argumentative corpora. We recommend that future research explores other languages, especially those underrepresented in argumentation mining.

Acknowledgments

We thank our colleagues in Innovations department of the Bundesdruckerei GmbH and the Hasso Plattner Institute for providing us with the opportunity to freely work on our research topics. Thank you for fostering an environment that encourages innovation and academic growth. The authors also sincerely thank the anonymous reviewers for their thoughtful recommendations that significantly improved this paper.

References

Mohammad Yeghaneh Abkenar and Manfred Stede. 2024. Neural mining of persian short argumentative texts. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 30–35.

Mohammad Yeghaneh Abkenar, Manfred Stede, and Stephan Oepen. 2021. Neural argumentation mining on essays and microtexts with contextualized word embeddings.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Basit Ali, Sachin Pawar, Girish Palshikar, and Rituraj Singh. 2022. Constructing a dataset of support and attack relations in legal arguments in court judgements using linguistic rules. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 491–500.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project debater apis: Decomposing the ai grand challenge. *arXiv preprint arXiv:2110.01029*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

Adrian de Wynter and Tommy Yuan. 2023. I wish to have an argument: Argumentative reasoning in large language models. *arXiv preprint arXiv:2309.16938*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024b. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-lingual argumentation mining for russian texts. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 134–144. Springer.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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*. Columbia Univ., New York, NY (United States).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv: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, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

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

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024a. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024b. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18608–18616.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the german online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153.

Ivan Namor and Manfred Stede. 2019. Mining italian short argumentative texts. In *Proceedings of the 5th Workshop on Argument Mining*.

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining. *arXiv preprint arXiv:1809.08145*.

Andreas Peldszus and Manfred Stede. 2015a. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815.

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

Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from english to portuguese. In *Proceedings of the 5th Workshop on Argument Mining, 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2024. Evaluating open language models across task types, application domains, and reasoning types: An in-depth experimental analysis. *arXiv preprint arXiv:2406.11402*.

Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

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*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, pages 1680–1691.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

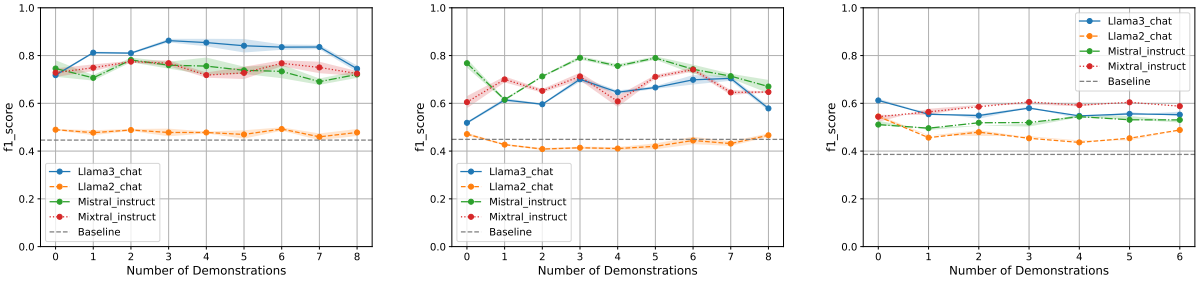
Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

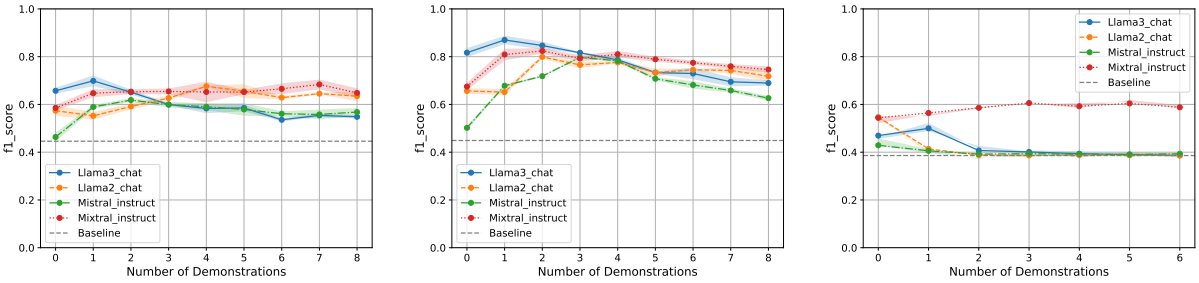
A Evaluation Results on Macro F1

B Prompt Design

In this section, we show four examples of our prompts designed for both vanilla and context-aware prompting methods.



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting

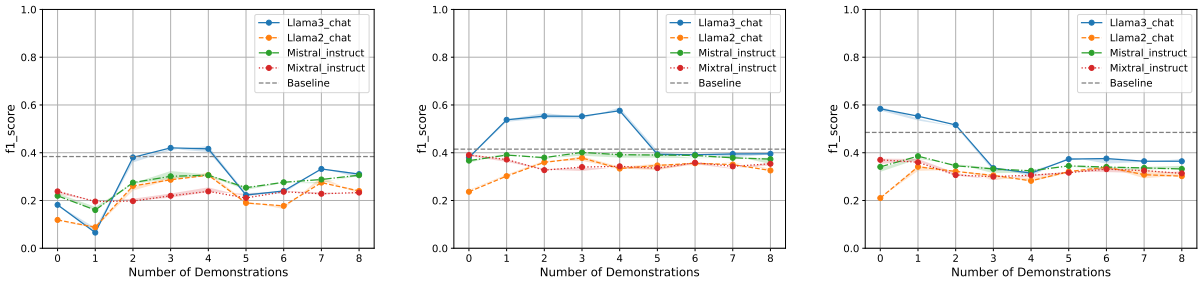


(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

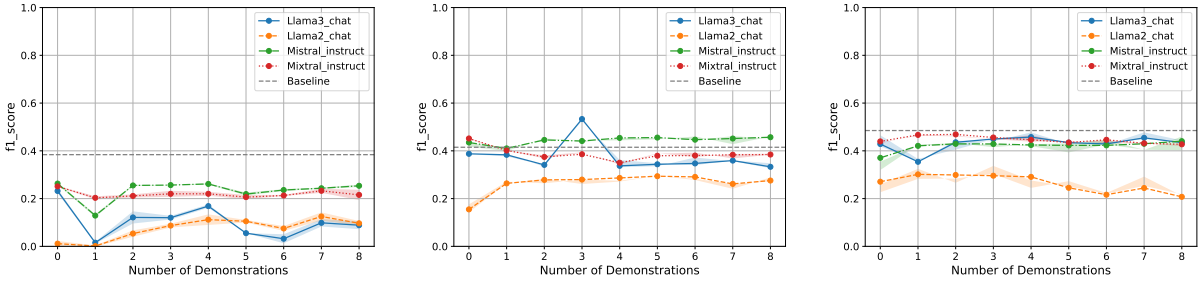
Figure 5: ACTC. The first row shows the model performance using vanilla prompting on three datasets whereas the below row shows the performance with the context-aware prompting.

System: You are an expert in linguistics and you are very good at argumentation mining. Now you are given a paragraph with indexs. Each sub-text is either the claim or premise. Your task is to find the claim in the paragraph. Provide the index of the claim in the text with < >. There is only one correct index.
Demo: Yes, it's annoying and cumbersome to separate your rubbish properly all the time. <2>Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. <3>But still Germany produces way too much rubbish, <4>and too many resources are lost when what actually should be separated and recycled is burnt. <5>We Berliners should take the chance and become pioneers in waste separation! The answer is: <5> One can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt. <2>And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles. <3>Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners. <4>Of course, first they'd actually need to be caught in the act by public order officers, <5>but once they have to dig into their pockets, their laziness will sure vanish! The answer is: <3> ...
Query: <1>For dog dirt left on the pavement dog owners should by all means pay a bit more. <2>Indeed it's not the fault of the animals, <3>but once you step in it, their excrement seems to stick rather persistently to your soles. The answer is:

Table 3: Example of Vanilla Prompting for ACTC task using AMT1 dataset.



(a) AMT1 vanilla prompting (b) AMT2 vanilla prompting (c) PE vanilla prompting



(d) AMT1 context-aware prompting (e) AMT2 context-aware prompting (f) PE context-aware prompting

Figure 6: ARC. The first row shows the model performance using vanilla prompting on three datasets where the below row shows the performance with the context-aware prompting.

System: You are an expert in linguistics and you are very good at argumentation Mining. Now you are given a sentence and a paragraph containing this sentence as a reference. Your task is to classify the sentence as either a Claim or a Premise according to the paragraph. Answer with <0> for Premise and <1> for Claim. There is only one Claim in the paragraph.

Demo: Please classify the sentence: Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. as either <1> for Claim or <0> for Premise in the given context: Yes, it's annoying and cumbersome to separate your rubbish properly all the time. Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. But still Germany produces way too much rubbish and too many resources are lost when what actually should be separated and recycled is burnt. We Berliners should take the chance and become pioneers in waste separation!
The answer is: <0>

Please classify the sentence: And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles. as either <1> for Claim or <0> for Premise in the given context: One can hardly move in Friedrichshain or Neuk0̈0f6lln these days without permanently scanning the ground for dog dirt. And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles. Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners. Of course, first they'd actually need to be caught in the act by public order officers, but once they have to dig into their pockets, their laziness will sure vanish!
The answer is: <0>
...

Query: Please classify the sentence: For dog dirt left on the pavement dog owners should by all means pay a bit more. as either <1> for Claim or <0> for Premise in the given context: For dog dirt left on the pavement dog owners should by all means pay a bit more. Indeed it's not the fault of the animals, but once you step in it, their excrement seems to stick rather persistently to your soles.
The answer is:

Table 4: Example of Context-aware Prompting for ACTC task using AMT1 dataset.

System: You are an expert in linguistics and you are very good at Relation Mining. Now you are given two sentences in an essay. Your task is to classify the relationship between the two sentences as 'Support' if Sentence 1 supports the stance of Sentence 2; or 'Attack' if Sentence 1 does not support Sentence 2. Provide only one word. DO NOT give explanation

Demo: Sentence 1:One who is living overseas will of course struggle with loneliness, living away from family and friends. Sentence 2:living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet.
The answer is: Attack

Sentence 1:What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others. Sentence 2:through cooperation, children can learn about interpersonal skills which are significant in the future life of all students.
The answer is: Support
...

Query: Sentence 1:it also has to be affordable for the consumer. Sentence 2:When a product is commonly used, it becomes trustworthy for the society, no matter what quality it is.
The answer is:

Table 5: Example of Vanilla Prompting for ARC task using PE dataset.

System: You are an expert in linguistics and you are very good at Relation Mining. Now you are given two sentences in an essay. Your task is to classify the relationship between the two sentences as 'Support' if Sentence 1 supports the stance of Sentence 2; or 'Attack' if Sentence 1 does not support Sentence 2. Use the context as supporting context. Provide only one word. DO NOT give explanation.

Demo: Sentence 1:One who is living overseas will of course struggle with loneliness, living away from family and friends. Sentence 2:living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet. Please classify the relationship as either Attack or Support based on the given context: Living and studying overseas It is every student's desire to study at a good university and experience a new environment. While some students study and live overseas to achieve this, some prefer to study home because of the difficulties of living and studying overseas. In my opinion, one who studies overseas will gain many skills throughout this experience for several reasons. First, studying at an overseas university gives individuals the opportunity to improve social skills by interacting and communicating with students from different origins and cultures. Compared to . . . . . in general life.
The answer is: Attack

Sentence 1:What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others. Sentence 2:through cooperation, children can learn about interpersonal skills which are significant in the future life of all students. Please classify the relationship as either Attack or Support based on the given context: Should students be taught to compete or to cooperate?It is always said that competition can effectively promote the development of economy. In order to survive in the competition, companies continue to improve their products and service, and as a result, the whole society prospers. However, when we discuss the issue of competition or cooperation, ... in one's success.
The answer is: Support
. . .

Query: Sentence 1:it is necessary to make sure that people can live a long life. Sentence 2:animal experiments have negative impact on the natural balance. Please classify the relationship as either Attack or Support based on the given context: Using animals for the benefit of the human beings with the rapid development of the standard of people's life, increasing numbers of animal experiments are done, new medicines and foods, for instance. Some opponents says that it is cruel to animals and nature, however, I believe that no sensible person will deny that it is a dramatically cruel activity to humanity if the latest foods or medicines are allowed to be sold without testing on animals. In my essay, I will discuss this issue from twofold aspects.First of all, as we all know, animals are friendly and vital for people, because if there are no animals in the world, the balance of nature will break down, and we, human, will die out as well. The animal experiments accelerate the vanishing of some categories of animals. In other words, doing this various testing is a hazard of human's future and next generation.Though animal experiments have negative impact on the natural balance, it is necessary to make sure that people can live a long life. To begin with, it is indisputable that every new kind food or pill may be noxious, and scientists must do something to insure that the new invention benefits people instead of making people ill or even dying. The new foods or medicines are invented to promote the quantity of human's life. Thus even if they are volunteers; they cannot take the place of animals to test the new foods or medicines. Furthermore, it also have potentially harm for human's health without any testing. To sum up, I reaffirm that although there is some disadvantages of animals' profits, the merits of animal experiments still outweigh the demerits.
The answer is:

Table 6: Example of Context-aware Prompting for ARC task using PE dataset.