Exploring the Potential of Large Language Models in Computational Argumentation

Guizhen Chen[*] [1,2] **Liying Cheng**[*†] [1,3] **Luu Anh Tuan**[2] **Lidong Bing**[1,3]
[1]DAMO Academy, Alibaba Group, Singapore [2]Nanyang Technological University, Singapore
[3]Hupan Lab, 310023, Hangzhou, China
{guizhen.chen, liying.cheng, l.bing}@alibaba-inc.com
{guizhen001, anhtuan.luu}@ntu.edu.sg

Abstract

Computational argumentation has become an essential tool in various domains, including law, public policy, and artificial intelligence. It is an emerging research field in natural language processing that attracts increasing attention. Research on computational argumentation mainly involves two types of tasks: argument mining and argument generation. As large language models (LLMs) have demonstrated impressive capabilities in understanding context and generating natural language, it is worthwhile to evaluate the performance of LLMs on diverse computational argumentation tasks. This work aims to embark on an assessment of LLMs, such as ChatGPT, Flan models, and LLaMA2 models, in both zero-shot and few-shot settings. We organize existing tasks into six main categories and standardize the format of fourteen openly available datasets. In addition, we present a new benchmark dataset on counter speech generation that aims to holistically evaluate the end-to-end performance of LLMs on argument mining and argument generation. Extensive experiments show that LLMs exhibit commendable performance across most of the datasets, demonstrating their capabilities in the field of argumentation. Our analysis offers valuable suggestions for evaluating computational argumentation and its integration with LLMs in future research endeavors. [1]

1 Introduction

Argumentation is a powerful and indispensable tool in various domains such as legality (Mochales and Moens, 2011; Grabmair et al., 2015), debating (Slonim et al., 2021; Li et al., 2020), and education (Stab and Gurevych, 2016). It plays a vital role in facilitating understanding between individuals by providing insights into different perspectives and their underlying reasons. Additionally, argumentation serves as a means of communicating convincing opinions, enhancing the acceptability of positions among readers. As computational argumentation becomes a growing research field in natural language processing (NLP) (Dietz et al., 2021; Habernal and Gurevych, 2016a; Atkinson et al., 2017; Wachsmuth et al., 2017; Holtermann et al., 2022; Barrow et al., 2021), researchers have dedicated considerable efforts to two distinct directions (Chakrabarty et al., 2019; Cheng et al., 2021; Alshomary et al., 2021; Bilu et al., 2019). The first direction, argument mining, focuses on understanding unstructured texts and automatically extracting various argumentative elements (Cabrio and Villata, 2018; Levy et al., 2014a; Rinott et al., 2015; Cheng et al., 2022). The other direction is argument generation, which aims to generate argumentative texts based on external knowledge (Hua et al., 2019; Schiller et al., 2020) or summarize key argument points. (Syed et al., 2021; Roush and Balaji, 2020).

Unlike classical structure prediction NLP tasks like named entity recognition that typically take a single sentence as the input and extract token-level information, computational argumentation tasks require discourse-level comprehension. This requirement makes it challenging and laborious to gather a large volume of labeled data for training, hindering the progress of research in this field. Fortunately, recent studies have shown that large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Tay et al., 2022; Touvron et al., 2023a) have demonstrated impressive performance on a wide variety of NLP tasks (Zhong et al., 2023; Pan et al., 2023b; Wang et al., 2023b; Cheng et al., 2023; Shen et al., 2023) in both zero-shot and few-shot settings. Given their strong capability in understanding long contexts and generating natural language, it is ex-

[*] Equal contribution. Guizhen Chen is under the Joint PhD Program between Alibaba and Nanyang Technological University.

[†] Liying Cheng is the corresponding author.

[1] Our data and code implementation are released at https://github.com/DAMO-NLP-SG/LLM-argumentation.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2309–2330
August 11-16, 2024 ©2024 Association for Computational Linguistics

citing yet still questionable how well LLMs can perform computational argumentation tasks without any supervised training.

In light of this, our objective is to investigate the performance of LLMs on diverse computational argumentation tasks. There are two main issues we aim to address in our study. Firstly, although there are existing surveys about argument mining (Peldszus and Stede, 2013), the systematic study of the broader definition of computational argumentation including argument mining and argument generation is under-explored. To bridge this gap, we categorize current computational argumentation tasks into two primary classes, comprising six distinct categories. In addition, we establish a standardized format and evaluation metrics for fourteen openly available datasets. Secondly, existing tasks and datasets either focus on argument mining or argument generation. To take a holistic approach, we propose a new task that integrates both argument mining and generation. This task is designed to generate counter speeches in response to debate speeches, which typically advocate a particular stance. We name them counter speech and supporting speech respectively in the remainder of our paper. This task requires the model to understand the argumentative structures in the supporting speech, meanwhile to generate the counter speech against the proposition. To facilitate the study, we construct a new document-to-document counterargument generation benchmark based on a debate database (Lavee et al., 2019).

To evaluate the performance of LLMs on computational argumentation tasks, we choose from both open-source and proprietary LLMs to conduct our main experiments, in zero-shot and few-shot settings. Our results reveal that LLMs exhibit promising performance in both argument mining and argument generation tasks. While LLMs might fail to achieve exceptionally high scores on specific metrics such as ROUGE, we hypothesize that the strict nature of these metrics could potentially underestimate the true potential of LLMs, which are inherently generative in nature. Human evaluation shows that LLMs are able to comprehend the core meaning of arguments and convey them effectively, even if the exact wording might not match. Collectively, these findings highlight the strengths of LLMs in grasping and effectively conveying the essence of arguments, showcasing their potential beyond what traditional metrics may suggest.

To summarize, our contributions include:

- We organize the existing computational argumentation tasks including argument mining and argument generation, and standardize the format of related datasets.
- We introduce a new task targeted at evaluating both argument mining and argument generation capabilities as a whole.
- To the best of our knowledge, we for the first time systematically evaluate the performance of multiple computational argumentation tasks using LLMs in zero-shot and few-shot settings.
- Extensive experimental results and analysis demonstrate the potential of LLMs in the computational argumentation research field and also suggest limitations in existing evaluation.

2 Background

Computational Argumentation Argumentation research has a long history (Walton et al., 2008; Hinton, 2019), aiming to persuade through logical propositions and achieve agreement among parties (Van Eemeren et al., 2004). Recently, computational argumentation has emerged as a significant field in NLP. The two main research directions are argument mining and argument generation, along with other directions such as persuasiveness of arguments (Habernal and Gurevych, 2016b) and quality assessment of arguments (Wachsmuth et al., 2017). Our work specifically focuses on argument mining and argument generation, where the detailed background can be found in Appendix A.

Large Language Models Recently, LLMs such as ChatGPT (OpenAI, 2023) have demonstrated strong capabilities in various NLP tasks. A surge of research has emerged to analyze and evaluate their performance on different types of tasks (Leiter et al., 2024; Liu et al., 2023d; Yang et al., 2024; Guo et al., 2023a; Laskar et al., 2023), including translation (Jiao et al., 2023; Liu et al., 2024), reasoning (Shakarian et al., 2023; Frieder et al., 2023; Liu et al., 2023b; Pan et al., 2023a), question answering (Tan et al., 2023; Pham et al., 2024), sentiment analysis (Zhong et al., 2023; Deng et al., 2023; Wang et al., 2023c; Zhang et al., 2023b; Nguyen et al., 2023a), text-to-SQL (Li et al., 2023; Liu et al., 2023a), dialogue understanding (Pan et al., 2023b; Fan and Jiang, 2023; Hu et al., 2023), relation extraction (Yuan et al., 2023; Nguyen et al., 2023b; Zheng et al., 2023), hate speech detection (Das et al., 2023; Hoang et al., 2024), summariza-

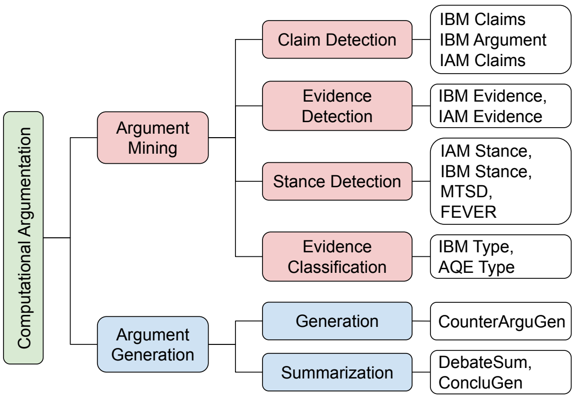

Figure 1: Explored tasks and datasets in this work.

tion (Nguyen and Luu, 2022; Yang et al., 2023; Wang et al., 2023a; Luo et al., 2023; Zhang et al., 2023a), and trustworthiness (Zhao et al., 2023, 2024) etc. However, it still lacks a systematic and thorough evaluation of computational argumentation using LLMs. Therefore, our work aims to explore the field of computational argumentation using LLMs by covering multiple tasks.

3 Tasks and Datasets

In this work, we systematically review existing tasks and datasets of computational argumentation and organize them in Figure 1. To maintain a balance for different tasks and datasets, we restrict our assessment by randomly sampling 500 examples from each dataset.

3.1 Argument Mining

We focus on the detection of argumentative components and their relations, which are the fundamental tasks in argument mining. We include a range of datasets with varying levels of difficulty, from simple binary tasks such as claim detection and evidence detection to harder ones like evidence type classification and stance detection. While we do not cover joint tasks such as end-to-end argument mining (Eger et al., 2017; Bao et al., 2022), these tasks could be transformed into a sequence of subtasks of identifying argumentative components and relations, which could be handled by our evaluated tasks.

Claim Detection A claim is a statement or proposition that asserts something to be true or false. In the context of argument mining, a claim is a key argument component that forms the basis of reasoning and debate. In claim detection tasks, the goal is to automatically extract claims from articles related to a specific debating topic (Levy et al., 2014b). We evaluate on datasets including IAM Claims (Cheng et al., 2022), IBM Claims (Levy et al., 2018a), and IBM Argument (Shnarch et al., 2020).

Evidence Detection Evidence is any information or data that supports or undermines a claim. In argument mining, evidence extraction involves automatically identifying and extracting relevant evidence from texts to substantiate claims (Rinott et al., 2015). Automating this process aids in comprehending and assessing arguments. By pinpointing relevant evidence, researchers can gain valuable insights into the underlying beliefs and motivations behind an argument. We evaluate evidence detection on the IBM Evidence dataset (Shnarch et al., 2018) and the IAM Evidence dataset (Cheng et al., 2022).

Stance Detection Stance represents a position towards a controversial topic, usually in the form of support and attack. Stance detection aims to determine whether a text supports, opposes, or remains neutral toward the topic. This task holds significance in domains such as politics (Habernal et al., 2017), fact-checking (Thorne et al., 2018; Guo et al., 2021), and journalism (Hanselowski et al., 2019), as it helps gauge public opinion and attitudes. Automated stance detection enhances the understanding and analysis of arguments across various applications. We use multiple datasets for evaluation, including FEVER (Thorne et al., 2018), IAM Stance (Cheng et al., 2022), IBM Stance (Levy et al., 2018b), and Multi-Target Stance Detection (MTSD) (Sobhani et al., 2017).

Evidence Type Classification Evidence type refers to the different categories of evidence that can be used to support or undermine a claim (Addawood and Bashir, 2016; Rinott et al., 2015). Examples of evidence types from previous works include statistics, expert opinions, facts, anecdotes, examples, etc. Automatic evidence type classification aids in understanding the strengths and weaknesses of an argument, particularly in fields such as debate, law, and policy. We use two datasets for evaluation, including IBM Type (Aharoni et al., 2014) and AQE Type (Guo et al., 2023b).

3.2 Argument Generation

We cover two main tasks: argument generation and argument summarization.

Generation Argument generation involves automatically generating arguments for or against a particular topic, to create persuasive and coherent arguments that can support or challenge a given position. We adopt the CounterArguGen dataset (Alshomary et al., 2021) for evaluation. There are two settings: generating a counter-argument given a claim with premises or generating based on a claim with weak premises.

Summarization The goal of argument summarization is to extract the main ideas and evidence supporting or challenging a particular claim or position and present them concisely and coherently. We evaluate two datasets: ConcluGen (Syed et al., 2021) and DebateSum (Roush and Balaji, 2020), which aim to summarize or give a conclusion for arguments. In the ConcluGen dataset, the corpus is augmented with three types of argumentative knowledge: topic, targets, and aspects. We study the effect of each argumentative knowledge and compare their respective performance with the base setting. In the DebateSum dataset, there are two settings. The abstractive summary generates a concise summary of the main points and arguments, while the extractive summary aims to extract relevant evidence from the passage to support the arguments.

3.3 Counter Speech Generation

Existing tasks in the field primarily center around either argument mining or argument generation. The former emphasizes language understanding, whereas the latter focuses on language generation. However, there is a lack of research comprehensively studying the overall argumentative capabilities of models. We contend that argument understanding and argument generation are two indispensable components of the broader computational argumentation landscape. Hence, a holistic perspective is necessary for evaluating the argumentative capabilities of models. Focusing solely on argument mining or argument generation provides only a partial understanding of their true potential.

In light of this, we propose a new task, counter speech generation, that aims to provide a more thorough evaluation of LLMs' argumentative capabilities. This task serves as a means to assess the model's capability to comprehend argumentative structures and generate counter-arguments accordingly. In debates, a supporting speech serves as a form of discourse intended to construct a specific idea or stance. It aims to provide compelling arguments and evidence in favor of a particular viewpoint. Counter speech generation, therefore, involves the task of generating a responsive or opposing speech in reaction to the supporting speech.

To the best of our knowledge, this is the first document-to-document counterargument generation task that simultaneously assesses a model from multiple perspectives including claim detection, stance detection, and argument generation. Earlier works focus on mining and retrieval of counterarguments (Wachsmuth et al., 2018; Bondarenko et al., 2020; Jo et al., 2021), which does not involve argument generation. Some focus on generating an opposing argument for a given statement which are typically short, informal texts from online forums (Alshomary et al., 2021; Hua and Wang, 2018; Hua et al., 2019). In contrast, ours consists of complete, formal speeches that are in the form of long argumentative texts, which potentially contain multiple arguments. Our task requires the model to first mine and analyze the main arguments in the original speech, then construct a complete and cohesive speech that addresses each key point. This expanded scope challenges the model to have a deeper understanding of argumentative structures from longer passages, while also requiring a heightened capacity to generate complete counter speeches.

To facilitate this study, we process a debate dataset (Lavee et al., 2019) by matching each supporting speech with the corresponding counter speech in a pool of debate scripts. We randomly sample 250 speech pairs for our zero-shot experiments. Given the constraint on limited annotated samples, we evaluate in a zero-shot setting only. Appendix B shows a data sample of this dataset.

4 Experiments

In this section, we discuss our choices of models, methods, and evaluation metrics.

4.1 Models

In our investigation, we examine the effectiveness of LLMs in directly performing inference on argument-related tasks without any fine-tuning. To accomplish this, we evaluate on open-source and proprietary models, including ChatGPT (`GPT-3.5-Turbo`) from OpenAI (OpenAI, 2023), `Flan-T5-XL`, `Flan-T5-XXL` (Chung et al., 2022) and `Flan-UL2` (Tay et al., 2022) from the Flan model family, as well as `Llama-2-7B`,

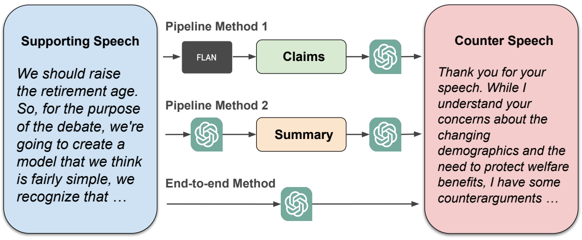
Supporting Speech
We should raise the retirement age. So, for the purpose of the debate, we're going to create a model that we think is fairly simple, we recognize that …
Pipeline Method 1
FLAN
Claims
Pipeline Method 2
Summary
End-to-end Method
Counter Speech
Thank you for your speech. While I understand your concerns about the changing demographics and the need to protect welfare benefits, I have some counterarguments …

Figure 2: Three different approaches for our proposed task on counter speech generation.

Llama-2-13B from the LLaMA2 series (Touvron et al., 2023b).

4.2 Methods

For tasks of a similar nature, we employ a consistent prompt format. More specifically, for argument mining tasks, we adhere to a standardized prompt template which consists of task definition and required output format. The task definition serves as a clear guideline for the LLMs to understand the task objective, while the required output format provides clarity on the expected output structure and restricts the generated response to a set of predefined labels to facilitate easier evaluation.

In contrast to argument mining tasks, output for argument generation tasks is more free-style and not constrained by any predetermined label space. The focus is on generating contextually relevant arguments. In order to tap into LLMs' linguistic knowledge and reasoning abilities, we adopt the prompts advised by ChatGPT. The prompt templates are available in Appendix C.

To tackle counter speech generation, we propose three different approaches[2], as shown in Figure 2. The first approach follows a pipeline method. We first identify the main claims from the supporting speech by determining if each sentence is a claim towards the given topic. We use Flan-T5-XXL due to its fast computation and strong capability in claim detection. After identifying all claims, we generate counterarguments that attack each claim detected in the supporting speech. For this step, GPT-3.5-Turbo is employed due to its strong generative ability.

Another pipeline approach is by generating a summary of the supporting speech. Initially, the key arguments in the supporting speech are summarized into a condensed representation of the main points. Subsequently, a counter speech is crafted to challenge these key arguments. In both steps, we use GPT-3.5-Turbo, which is adept at handling long inputs and comprehending long contexts.

Unlike the two-step approaches, the third method is a one-step process where we directly prompt GPT-3.5-Turbo to respond to the supporting speech by challenging the main arguments. This approach serves as a means of gauging the model's ability to internally identify key arguments and generate a respective counter speech.

4.3 Evaluation

To evaluate argument mining tasks, we use both accuracy and F1 score as the metrics.

To assess argument generation and counter speech generation, we employ a wide range of automatic evaluation metrics, including ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), METEOR (Denkowski and Lavie, 2011) and BERTScore (Zhang et al., 2019). The ROUGE scores assess the quality based on the overlap with the reference arguments, while METEOR also considers synonyms, paraphrases, and stemming. On the other hand, BERTScore takes into account the semantic context. We also conduct human evaluation to complement the results from automatic evaluation.

4.4 Previous SOTA

To compare our results against existing state-of-the-arts (SOTA), we either finetune pre-trained language models (PLMs) or leverage available checkpoints to conduct inference on our sampled test set. Training details are reported in Appendix D.

5 Results and Discussion

In this section, we discuss the main results and provide insights into the performance of various LLMs on argument mining, argument generation and counter speech generation.

5.1 Results on Argument Mining

Table 1 shows the zero-shot performance of three representative models, GPT-3.5-Turbo, Flan-UL2 and Llama-2-13B, across 11 argument mining datasets. Statistical tests[3] are conducted to show if the LLM's predictions are significantly different from the random observations. Results of other models including Flan-T5-XL, Flan-T5-XXL and Llama-2-7B are available in Appendix E.

[2] Note that other combinations of models could be used for each approach. Here we only employ the strongest model for each task/subtask, guided by the results in Section 5.

[3] We use McNemar's test (Mcnemar, 1947) following the guidelines by Dror et al. (2018).

Model	Claim Detection			Evidence Detection		Stance Detection				Evidence Classification	
	IBM Claims	IBM Argument	IAM Claims	IBM Evidence	IAM Evidence	IAM Stance	IBM Stance	MTSD	FEVER	IBM Type	AQE Type
					Acc.						
Random	50.20	48.00	52.00	49.20	51.00	45.80	47.00	10.00	29.40	33.20	19.20
GPT-3.5-Turbo	72.00	55.80	68.20	52.20	45.00	59.00	33.80	**41.00**	33.40	**73.40**	**58.20**
Flan-UL2 (20B)	**74.80**	**63.60**	**83.80**	**64.80**	**71.40**	**65.00**	**58.20**	15.40	**35.40**	68.60	21.60
Llama-2-13B	36.00	44.20	44.80	25.40	36.60	14.60	4.00	25.40	0.40	5.20	3.80
					F_1						
Random	55.53	52.01	64.28	49.62	58.05	51.76	50.05	12.59	**33.93**	33.51	24.42
GPT-3.5-Turbo	**72.19**	56.16	76.35	50.44	51.48	58.99	36.26	**42.27**	20.33	**72.39**	**59.95**
Flan-UL2 (20B)	71.80	**62.06**	**86.80**	**64.45**	**75.70**	**63.71**	**59.70**	13.38	28.06	67.34	15.68
Llama-2-13B	40.61	41.73	56.84	21.99	46.05	18.30	6.28	12.51	0.77	8.22	4.59
					p-value						
GPT-3.5-Turbo	2.69e-11*	1.61e-01	1.70e-08*	6.71e-01	1.71e-03*	1.27e-02*	2.25e-07*	0.00e+00*	5.83e-01	0.00e+00*	0.00e+00*
Flan-UL2 (20B)	5.86e-14*	1.04e-04*	0.00e+00*	2.05e-04*	2.67e-07*	9.58e-06*	1.35e-02*	1.00e+00	2.19e-01	0.00e+00*	1.88e-01
Llama-2-13B	1.79e-09*	1.29e-03*	4.50e-01	9.47e-03*	1.44e-11*	0.00e+00*	0.00e+00*	3.44e-07*	4.08e-01	0.00e+00*	6.99e-15*

Table 1: Zero-shot performance on argument mining tasks. Datasets with binary class are underlined green. Datasets involving multi-class are underlined red. Highest accuracy and F1 score for each task are in bold. $*$ indicates statistically significant results that are different from random observations with a significant level of $\alpha = 0.05$.

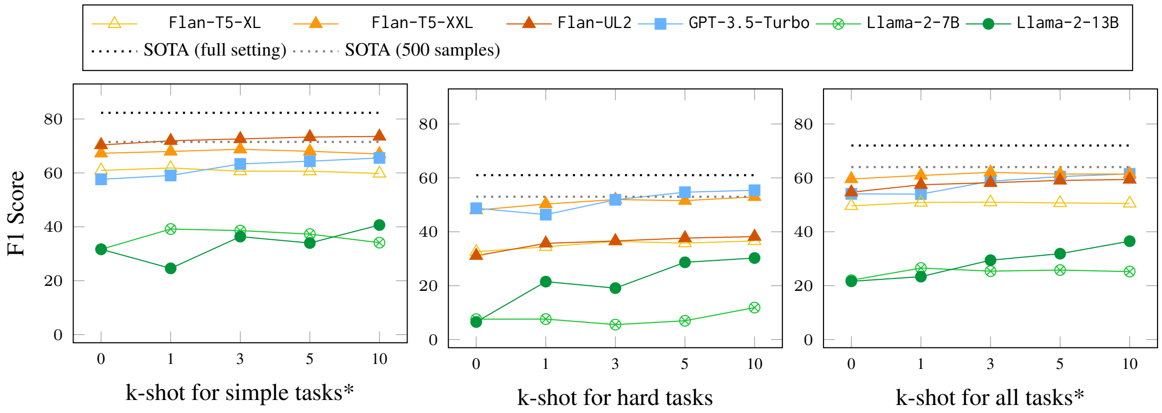

Figure 3: Few-shot performance comparison on argument mining tasks. Results of previous SOTA (using full setting and 500 samples) are also shown for easy comparison. *: Note that we exclude *IBM Argument* because the train set is smaller than 500.

To analyze, we categorize tasks into simple and hard tasks based on the number of classes involved. Binary classification tasks, including claim detection, evidence detection, IAM stance detection and IBM stance detection, are classified as simple tasks. Tasks with more than two labels, including evidence classification, FEVER stance detection and MTSD stance detection, are classified as hard tasks.

Overall, both GPT-3.5-Turbo and Flan-UL2 perform decently in the zero-shot setting, surpassing the random baseline with most results being statistically different. However, Llama-2-13B falls short of the random baseline in the majority of the tasks, notably in more challenging tasks like FEVER stance detection and evidence classification. This highlights its limitations in capturing nuanced stances and comprehending evidence types within the zero-shot context, which relies on sufficient prior knowledge in the model.

Comparing GPT-3.5-Turbo and Flan-UL2, Flan-UL2 consistently demonstrates higher proficiency in tasks like claim detection, evidence detection, and certain stance detection tasks that are mostly binary classification tasks. However, its performance diminishes when encountered with tasks that involve more than two classes, such as MTSD stance detection and AQE evidence classification. In contrast, GPT-3.5-Turbo generally demonstrates superior performance in these multi-class scenarios.

Figure 3 shows the effects of increasing shots on different models and task difficulties. In general, while there remain certain gaps between the few-shot performance of LLMs and finetuned PLMs using the full training set, it is worth noting a significant trend: by simply prompting LLMs with less than 10 demonstrations, they are able to close the gaps and match the performance of finetuned PLMs trained with 500 samples.

Comparing among models, we notice that the choice of model is crucial, as different models exhibit varying levels of proficiency across different

Task	Dataset	Setting	Method	k-shot	BERTScore	ROUGE1	ROUGE2	ROUGEL	METEOR
Generation	CounterArguGen	Premises	Alshomary et al. (2021)	-	82.60	17.76	1.36	10.66	14.85
			GPT-3.5-Turbo	k=0	83.50	18.36	1.58	11.07	17.60
		Weak Premises	Alshomary et al. (2021)	-	82.53	17.34	1.12	10.33	14.65
			GPT-3.5-Turbo	k=0	84.06	19.75	2.03	11.95	17.63
Summarization	ConcluGen	Base	Syed et al. (2021)	-	84.78	8.16	0.47	7.15	6.02
			GPT-3.5-Turbo	k=0	85.53	13.99	3.20	10.78	21.28
			GPT-3.5-Turbo	k=1	$86.51_{0.15}$	$16.80_{0.33}$	$3.86_{0.18}$	$12.96_{0.38}$	$20.34_{0.24}$
			GPT-3.5-Turbo	k=3	$86.95_{0.20}$	$18.54_{0.91}$	$4.66_{0.52}$	$14.53_{0.78}$	$21.27_{0.59}$
			GPT-3.5-Turbo	k=5	$87.19_{0.27}$	$19.39_{0.78}$	$5.09_{0.47}$	$15.21_{0.85}$	$21.50_{0.11}$
		Aspects	Syed et al. (2021)	-	89.32	31.47	16.90	28.94	27.61
			GPT-3.5-Turbo	k=0	85.47	13.79	3.25	10.43	21.70
			GPT-3.5-Turbo	k=1	$86.16_{0.25}$	$16.41_{0.92}$	$3.93_{0.32}$	$12.42_{0.73}$	$21.96_{0.81}$
			GPT-3.5-Turbo	k=3	$86.77_{0.09}$	$18.59_{0.77}$	$5.00_{0.43}$	$14.36_{0.53}$	$22.44_{0.53}$
			GPT-3.5-Turbo	k=5	$87.09_{0.15}$	$19.86_{0.70}$	$5.56_{0.45}$	$15.61_{0.73}$	$22.88_{0.36}$
		Targets	Syed et al. (2021)	-	89.18	30.58	15.73	27.71	26.28
			GPT-3.5-Turbo	k=0	85.68	14.69	3.61	11.03	22.17
			GPT-3.5-Turbo	k=1	$86.67_{0.19}$	$18.55_{0.74}$	$4.83_{0.55}$	$14.14_{0.67}$	$22.47_{0.87}$
			GPT-3.5-Turbo	k=3	$86.94_{0.27}$	$19.32_{1.17}$	$5.24_{0.72}$	$15.00_{0.96}$	$21.88_{0.77}$
			GPT-3.5-Turbo	k=5	$87.14_{0.32}$	$19.83_{1.22}$	$5.56_{0.61}$	$15.58_{0.95}$	$21.76_{0.78}$
		Topic	Syed et al. (2021)	-	89.38	32.34	17.42	29.45	28.22
			GPT-3.5-Turbo	k=0	85.75	15.08	3.53	11.35	22.31
			GPT-3.5-Turbo	k=1	$86.78_{0.52}$	$18.47_{1.59}$	$5.13_{1.22}$	$14.47_{1.92}$	$21.80_{0.93}$
			GPT-3.5-Turbo	k=3	$87.14_{0.18}$	$19.87_{1.17}$	$5.69_{0.86}$	$15.72_{0.99}$	$22.28_{1.43}$
			GPT-3.5-Turbo	k=5	$87.42_{0.22}$	$20.63_{1.35}$	$6.14_{1.00}$	$16.48_{1.11}$	$21.90_{1.47}$
	DebateSum	Abstractive	T5-base	-	82.88	11.39	1.65	10.41	6.00
			GPT-3.5-Turbo	k=0	84.25	10.35	2.06	8.28	16.25
			GPT-3.5-Turbo	k=1	$84.52_{0.24}$	$11.45_{0.71}$	$2.24_{0.24}$	$9.08_{0.60}$	$16.61_{0.50}$
			GPT-3.5-Turbo	k=3	$84.74_{0.20}$	$12.12_{0.55}$	$2.34_{0.19}$	$9.58_{0.51}$	$16.68_{0.20}$
			GPT-3.5-Turbo	k=5	$84.72_{0.17}$	$12.07_{0.42}$	$2.32_{0.06}$	$9.56_{0.35}$	$16.75_{0.05}$
		Extractive	Roush and Balaji (2020)	-	85.90	59.06	44.37	57.48	56.55
			GPT-3.5-Turbo	k=0	88.36	49.76	30.88	37.89	40.62
			GPT-3.5-Turbo	k=1	$88.84_{0.22}$	$51.91_{1.27}$	$34.51_{1.81}$	$41.19_{1.98}$	$41.85_{1.58}$
			GPT-3.5-Turbo	k=3	$89.52_{0.22}$	$55.33_{1.35}$	$41.05_{2.30}$	$46.80_{1.88}$	$47.07_{2.42}$
			GPT-3.5-Turbo	k=5	$89.43_{0.16}$	$54.99_{0.87}$	$40.09_{1.88}$	$45.95_{1.44}$	$46.02_{1.52}$

Table 2: Performance of GPT-3.5-Turbo on argument generation tasks. The results are averaged over 3 random seeds for all few-shot experiments.

tasks. While Flan models excel in simple tasks, the performances of Flan-T5-XL and Flan-UL2 lag behind that of GPT-3.5-Turbo in hard tasks, even with an increased number of shots. Overall, Flan-T5-XXL appears to be most robust, consistently demonstrating strong performance across diverse tasks.

Secondly, larger models are not necessarily superior to smaller models. Upon comparing the two LLaMA models, Llama-2-13B generally outperforms Llama-2-7B. However, one interesting exception surfaces when the input is minimal. Surprisingly, Llama-2-7B proves to be more effective than Llama-2-13B in simple tasks. For Flan models, larger models consistently outperform their smaller counterparts in simple tasks. This trend, however, does not hold in the case of more challenging tasks. Notably, Flan-T5-XL (3B) model performs comparably to Flan-UL2 (20B) in hard tasks, despite its significantly smaller size. Furthermore, the 11B Flan-T5-XXL model showcases remarkable performance, even though it is smaller than both Flan-UL2 and Llama-2-13B. This suggests that, for certain complex tasks, the performance of the model may not be solely determined by its size.

Furthermore, increasing demonstrations have varying effects on different models. GPT-3.5-Turbo generally benefits from more shots. For Flan models, the gain in performance is not obvious. Llama models, on the other hand, exhibit mixed performance in response to more demonstrations. The larger model demonstrates notable performance improvement, particularly in hard tasks, when provided with more shots. However, the smaller model does not exhibit performance gain from additional demonstrations. In fact, when it comes to simple tasks, providing more shots has a negative impact. It appears that longer contexts might introduce noise or unnecessary information that could potentially hinder the performance of smaller models.

5.2 Results on Argument Generation

Table 2 presents the performance of GPT-3.5-Turbo on argument generation tasks.

Compared to existing SOTA, `GPT-3.5-Turbo` already outperforms previous methods in several tasks including CounterArguGen, ConcluGen in the base setting, as well as abstractive summarization in DebateSum.

Although previous methods excel in other ConcluGen settings, we attribute their high performance to additional annotations encoding specific aspects, targets, or topics. Such manual annotations are task-specific and extremely costly. `GPT-3.5-Turbo`, on the other hand, achieves comparable results across different settings regardless of the presence of encoded information. Furthermore, the contrasting results from different evaluation metrics reveal an interesting pattern: the ROUGE scores are generally low but the BERTScores are high. The low ROUGE scores indicate that there are only a few overlaps between the generated text and the reference text. The high BERTScore indicates that the semantic meaning of the generated text is highly similar to the reference text. This suggests that although the generated text may not match the reference text in terms of exact wording or specific phrases, it successfully captures the underlying semantic meaning. To further support this, we provide several illustrative examples in Appendix F. Both automatic evaluation and quality analysis show that `GPT-3.5-Turbo` grasps the essence of the content and conveys it effectively, even if the choice of words or phrasing differs from the reference.

For extractive summarization, previous method (Roush and Balaji, 2020) relies on word-level classification, wherein each word is predicted as either "underlined" or "not-underlined", which is inefficient and compromises the coherence of the generated sentences. In contrast, `GPT-3.5-Turbo` avoids the high training cost and generates coherent sentences. Additionally, our quality analysis shows that `GPT-3.5-Turbo` is able to identify important information accurately. Case studies can be found in Appendix G.

In addition, we notice that `GPT-3.5-Turbo` exhibits incremental performance improvements as the number of shots increases. The performance gains are relatively modest compared to those observed in argument mining tasks. This implies that `GPT-3.5-Turbo` is inherently proficient in argument generation without necessitating more demonstrations.

We also evaluate other models including `Llama-2-7B`, `Llama-2-13B` and `Flan-UL2` which could accommodate long context. Performance of other models are available in Appendix H. All models exhibit similar trends with the above except `Flan-UL2` - its advantage in extractive summarization is less apparent compared to the other models.

Method	**BERTScore**	**ROUGE1**	**ROUGE2**	**ROUGEL**	**METEOR**
Pipeline (Claims)	$80.33_{0.08}$	$31.00_{0.54}$	$3.70_{0.24}$	$13.26_{0.14}$	$21.92_{1.21}$
Pipeline (Summary)	$82.23_{0.06}$	$23.73_{6.19}$	$4.28_{0.92}$	$11.60_{2.06}$	$10.89_{3.41}$
End-to-end	$82.51_{0.05}$	$30.10_{1.08}$	$5.70_{0.18}$	$13.65_{0.22}$	$14.48_{0.77}$

Table 3: Automatic evaluation results of counter speech generation. The average scores are calculated based on three distinct sets of prompts to account for the potential sensitivity of zero-shot performance to prompt designs.

5.3 Results on Counter Speech Generation

Automatic Evaluation Table 3 shows the results from automatic evaluation. The end-to-end approach surpasses the summarization pipeline approach across all metrics. This highlights the model's strong capability of internalizing and synthesizing information from the supporting speech without the need for intermediate steps.

Comparing the end-to-end approach to the claim detection pipeline approach, the former lags behind in ROUGE-1 and METEOR, but surpasses in BERTScore, ROUGE-2, and ROUGE-L. To determine which approach is superior, we conduct human evaluation for a more complete understanding of the performance of these two approaches.

Human Evaluation We hire 2 human judges who are professional English speakers to manually evaluate the quality of counter speeches generated by the claim pipeline approach and the end-to-end approach on 50 random samples. For each test instance, we provide the judges with supporting speeches along with randomly ordered counter speeches from the two methods, and ask the judges to individually evaluate the generation quality based on the following criteria:

- *Fluency (Flu.)*: Is the generation fluent, grammatical, and without unnecessary repetitions?
- *Persuasiveness (Per.)*: Is the text able to convince you to adopt a certain belief or attitude?
- *% of arguments addressed (% Arg.)*: Does the counter speech address all claims/arguments in the supporting speech?

Fluency and persuasiveness, graded on a scale of 1 to 5, are to assess model's argument generation capability. To evaluate the model's argument mining

Method	*Flu.*	*Per.*	*% Arg.*
Pipeline (Claims)	3.56	2.8	78%
End-to-end	4.32	3.8	95%

Table 4: Human evaluation scores on 50 test samples.

ability, we use % of arguments addressed, calculated by the number of addressed arguments in the counter speech over the total number of arguments in the supporting speech. This metric reflects how effectively the model is able to identify arguments, either explicitly in the two-step approach or implicitly in the one-step approach.

In Table 4, it is evident that the end-to-end approach outperforms the pipeline approach on all 3 metrics. In specific, the pipeline approach is not able to address as many arguments as the end-to-end approach, possibly due to the potential loss of information during the intermediate step. In the pipeline method, information from the supporting speech undergoes processing, such as summarization or claim detection, before the final counter speech is generated. This might result in information loss or distortion, which could negatively impact the overall coherence and effectiveness of the generated response, which in turn affects the fluency and persuasiveness scores. In contrast, the one-step approach bypasses the intermediate stage, allowing the model to directly engage with the supporting speech and generate a counter speech in a more holistic manner. We show qualitatively in Appendix I.

6 Conclusion and Broader Impacts

In this paper, we have made several significant contributions to the field of computational argumentation research.

Firstly, our efforts in organizing the diverse landscape of argumentation-related tasks and standardizing the format of related datasets are crucial for future research in designing domain-specific large-scale models for argumentation.

Secondly, we for the first time systematically evaluate the performance of multiple computational argumentation tasks using LLMs in zero-shot and few-shot settings. Traditional approaches for computational argumentation rely heavily on supervised fine-tuning that requires a large amount of labeled data, hindering the progress of research in this field. Our exploration of low-resource settings addresses a gap in previous computational argumentation research, demonstrating the potential of LLMs in scenarios with limited training data.

Furthermore, we introduce a new counter speech generation benchmark that evaluates models' capability in both argument mining and argument generation. Our extensive experimental results and analysis demonstrate the potential of LLMs in computational argumentation, while also highlighting existing limitations in evaluating computational argumentation tasks.

Overall, our paper provides important insights and valuable resources for researchers interested in the field of computational argumentation, which will potentially inspire further advancement in this exciting area.

Limitation

In the field of computational argumentation, there are more tasks involved. In this work, we only cover argument mining and argument generation, as these two categories are the most fundamental in this field. By understanding and establishing the performance of these core tasks, we could progress to tackle other tasks in our future study. In addition, it is challenging and laborious to conduct human evaluation on the full argument generation datasets. To address this, we could use GPT-4 as an evaluator (Liu et al., 2023c).

Acknowledgments

This work was supported by DAMO Academy through DAMO Academy Research Intern Program. We would like to thank Wenxuan Zhang for providing feedback on the paper.

References

Aseel Addawood and Masooda N. Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *ArgMining@ACL*.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2021. Argument undermining: Counter-argument generation by attacking weak premises. In *Proceedings of ACL-IJCNLP*.

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo R. Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI Mag.*, 38:25–36.

Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. 2022. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10437–10449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joe Barrow, R. Jain, Nedim Lipka, Franck Dernoncourt, Vlad I. Morariu, Varun Manjunatha, Douglas W. Oard, Philip Resnik, and Henning Wachsmuth. 2021. Syntopical graphs for computational argumentation tasks. In *Proceedings of ACL*.

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of ACL*.

Alexander Bondarenko, Matthias Hagen, Martin Potthast, Henning Wachsmuth, Meriem Beloucif, Chris Biemann, Alexander Panchenko, and Benno Stein. 2020. Touché: First shared task on argument retrieval. *Advances in Information Retrieval*, 12036:517 – 523.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of IJCAI*.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen McKeown, and Alyssa Hwang. 2019. Ampersand: Argument mining for persuasive online discussions. In *Proceedings of EMNLP-IJCNLP*.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. Iam: A comprehensive and large-scale dataset for integrated argument mining tasks. In *Proceedings of ACL*.

Liying Cheng, Xingxuan Li, and Lidong Bing. 2023. Is gpt-4 a good data analyst? In *Findings of EMNLP*.

Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument pair extraction via attention-guided multi-layer multi-cross encoding. In *Proceedings of ACL-IJCLP*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating chatgpt's performance for multilingual and emoji-based hate speech detection. *ArXiv*, abs/2305.13276.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. Llms to the moon? reddit market sentiment analysis with large language models. *Companion Proceedings of the ACM Web Conference 2023*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Emmanuelle-Anna Dietz, Antonis C. Kakas, and Loizos Michael. 2021. Computational argumentation & cognitive ai. In *International Conference on Advances in Computing and Artificial Intelligence*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Yaxin Fan and Feng Jiang. 2023. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. *ArXiv*, abs/2305.08391.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and J J Berner. 2023. Mathematical capabilities of chatgpt. *ArXiv*, abs/2301.13867.

Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. 2015. Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. In *Proceedings of international conference on artificial intelligence and law*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023a. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *ArXiv*, abs/2301.07597.

Jia Guo, Liying Cheng, Wenxuan Zhang, Stanley Kok, Xin Li, and Lidong Bing. 2023b. Aqe: Argument quadruplet extraction via a quad-tagging augmented generative approach. In *Findings of ACL*.

Zhijiang Guo, M. Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking. *TACL*, 10:178–206.

Ivan Habernal and Iryna Gurevych. 2016a. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43:125–179.

Ivan Habernal and Iryna Gurevych. 2016b. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of EMNLP*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of NAACL*.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. *ArXiv*, abs/1911.01214.

Martin Hinton. 2019. Language and argument: a review of the field. *Research in Language*, 17:103 – 93.

Nhat M Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. Toxcl: A unified framework for toxic speech detection and explanation. *arXiv preprint arXiv:2403.16685*.

Carolin Holtermann, Anne Lauscher, and Simone Paolo Ponzetto. 2022. Fair and argumentative language modeling for computational argumentation. In *Proceedings of ACL*.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3953–3957.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of ACL*, pages 219–230.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *ArXiv*, abs/2301.08745.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of EMNLP*, pages 3074–3094.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Tamar Lavee, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Shachar Mirkin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. In *ArgMining@ACL*.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2024. Chatgpt: A meta-analysis after

2.5 months. *Machine Learning with Applications*, 16:100541.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014a. Context dependent claim detection. In *Proceedings of COLING*.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014b. Context dependent claim detection. In *Proceedings of ICCL*.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018a. Towards an argumentative content search engine using weak supervision. In *Proceedings of ICCL*, pages 2066–2081.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018b. Towards an argumentative content search engine using weak supervision. In *Proceedings of ICCL*.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of EMNLP*.

Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Chenhao Ma, Kevin C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *ArXiv*, abs/2305.03111.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL*.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023a. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *ArXiv*, abs/2303.13547.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yuexin Zhang. 2023b. Evaluating the logical reasoning ability of chatgpt and gpt-4. *ArXiv*, abs/2304.03439.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023d. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *ArXiv*, abs/2303.15621.

Quinn Mcnemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*.

Cong-Duy Nguyen, Thong Nguyen, Duc Vu, and Anh Luu. 2023a. Improving multimodal sentiment analysis: Supervised angular margin-based contrastive learning for enhanced fusion representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14714–14724.

Huy Nguyen, Chien Nguyen, Linh Ngo, Anh Luu, and Thien Nguyen. 2023b. A spectral viewpoint on continual relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9621–9629.

Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11103–11111.

OpenAI. 2023. Introducing chatgpt.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023a. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023b. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *ArXiv*, abs/2304.04256.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Informatics Nat. Intell.*, 7:1–31.

Phuoc Van Long Pham, Anh Vu Duc, Nhat Minh Hoang, Xuan Long Do, and Anh Tuan Luu. 2024. Chatgpt as a math questioner? evaluating chatgpt on generating pre-university math questions. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 65–73.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of EMNLP*.

Allen Roush and Arvind Balaji. 2020. Debatesum: A large-scale argument mining and summarization dataset. *ArXiv*, abs/2011.07251.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Aspect-controlled neural argument generation. In *Proceedings of NAACL*.

Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An independent evaluation of chatgpt on mathematical word problems (MWP). In *Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023*, volume 3433 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of EMNLP*.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of ACL*.

Eyal Shnarch, Leshem Choshen, Guy Moshkowich, Noam Slonim, and Ranit Aharonov. 2020. Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of EMNLP*.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of EACL*.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43:619–659.

Shahbaz Syed, Khalid Al-Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of ACL*.

Yiming Tan, Dehai Min, Y. Li, Wenbo Li, Na Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *ArXiv*, abs/2303.07992.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier García, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2022. Ul2: Unifying language learning paradigms. In *Proceedings of ICLR*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

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

Frans H Van Eemeren, Robert Grootendorst, and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of EACL*.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of ACL*.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Cross-lingual summarization via chatgpt. *ArXiv*, abs/2302.14229.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023c. Is chatgpt a good sentiment analyzer? a preliminary study. *ArXiv*, abs/2304.04339.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*. Just Accepted.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *ArXiv*, abs/2302.08081.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *ArXiv*, abs/2304.05454.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via chatgpt for faithful summary generation. *ArXiv*, abs/2304.04193.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Wenxuan Zhang, Yue Deng, Bing-Quan Liu, Sinno Jialin Pan, and Lidong Bing. 2023b. Sentiment analysis in the era of large language models: A reality check. *ArXiv*, abs/2305.15005.

Shuai Zhao, Leilei Gan, Luu Anh Tuan, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. 2024. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.12168*.

Shuai Zhao, Jinming Wen, Luu Anh Tuan, Junbo Zhao, and Jie Fu. 2023. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. *arXiv preprint arXiv:2305.01219*.

Yandan Zheng, Anran Hao, and Anh Tuan Luu. 2023. Jointprop: joint semi-supervised learning for entity and relation extraction with heterogeneous graph-based propagation. *arXiv preprint arXiv:2305.15872*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *ArXiv*, abs/2302.10198.

A More Background

A.1 Argument Mining

Argument mining is a rapidly emerging field of NLP that aims to automatically identify and extract arguments and their components from textual data. With the increasing volume of digital text available online, the need for automated methods to analyze and understand arguments has become more pressing. By identifying the arguments in natural language text, researchers can better understand the underlying beliefs, values, and motivations that drive human behavior. As such, argument mining is a core task of research within NLP that is poised to make significant contributions to a wide range of fields.

A.2 Argument Generation

With the understanding of the argumentative structures within the text through argument mining, the next step is to explore how to generate arguments. Argument generation and argument summarization are two related tasks within computational argumentation that have the potential to transform the way we create and consume arguments. Argument generation involves the automatic creation of persuasive text, such as generating a sentence attacking another standpoint, that can be used to influence a group of readers. Argument summarization, on the other hand, involves the automatic summarization of arguments, enabling users to quickly and easily understand complex arguments without having to read through lengthy documents. For example, in the law domain, large amounts of legal documents need to be analyzed and understood in a time-sensitive manner. As such, argument generation and summarization are two key areas of research within NLP that have the potential to significantly streamline the process of argumentation in various domains.

B Data Sample on Counter Speech Generation

Table 5 shows a data sample from our benchmark dataset for the proposed counter speech generation task. The topic is "Nationalism does more harm than good". The supporting speech is the input, and the counter speech written by humans is considered the output.

C Prompt Templates

C.1 Prompt Templates for Argument Mining Tasks

Table 6 shows the prompt templates for selected argument mining tasks, including claim detection and stance detection.

C.2 Prompt Templates of Argument Generation Tasks

Table 7 shows the prompt templates for argument generation tasks, including counter argument generation and abstractive summarization.

D Training Details of SOTA

For argument mining tasks, we train sentence-pair classifiers based on pre-trained models such as BERT (Devlin et al., 2019) following the settings reported by Cheng et al. (2022). Dataset statistics can be found in Table 8. For datasets where training sets are not available, we randomly sample 500 data points to reserve for the test set and make use of all the remaining for training.

For CounterArguGen, we directly evaluate based on the released predictions (Alshomary et al., 2021).

For ConcluGen (Syed et al., 2021), we use the available checkpoints and conduct inference on our sampled test set. The released checkpoints were trained on the reported training sets, while our 500 samples were sampled from the test sets for proper train-test split.

For DebateSum tasks, we randomly sample 90000 data for train set, 10000 for development set, and 500 for test set, since the original train test split is not specified. Specifically, for abstractive summarization, we finetune a T5-base (Raffel et al., 2020), a popular and performant generative model, using the AdamW optimizer with a learning rate of 1e-4, a fixed batch size of 4, and 3 training epochs. For DebateSum extractive summarization, we follow the settings reported by Roush and Balaji (2020).

E Additional Results on Argument Mining

Table 9 shows the zero-shot performance of `Flan-T5-XL`, `Flan-T5-XXL` and `Llama-2-7B` on argument mining tasks.

F Quality Analysis on Argument Generation Tasks

To further support our claims in Section 5.2, we show three examples of references and predictions from the ConcluGen dataset in Table 10. For instance, in the third pair, while the generated text uses “should” instead of “need to”, and “be held to the same standards” instead of “follow the same set of rules”, it effectively conveys the same meaning as the reference. These observations imply that the generated text might have used different wordings but the overall semantic meaning is similar to that of the reference text, which further supports our claims.

G Case Study on Extractive Summarization

To further support our claims in Section 5.2, we show 2 examples in Table 11 and 12. It can be observed that Longformer tends to generate incoherent sentences, while `GPT-3.5-Turbo` can generate coherent sentences and can extract important information from the input.

H Additional Results on Argument Generation

Table 13, 14 and 15 display the evaluation results of argument generation tasks using `Llama-2-7B`, `Llama-2-13B` and `Flan-UL2` respectively.

I Case Study on Counter Speech Generation

Table 16 shows a case study of the data sample shown in Table 5 for the proposed counter speech generation task. The pipeline approach by extracting claims first tends to generate repetitive phrases, and does not attack all the claims stated in the supporting speech. In contrast, the end-to-end approach is more concise and attacks the claims in the supporting speech.

Topic	Nationalism does more harm than good
Supporting Speech	Nationalism does more harm than good. What's important to recognize about nationalism right at the outset is that it doesn't arise from anything natural about the peoples that express nationalist attitudes. There's nothing about german nationalists or french nationalists or chinese nationalists that makes those types of groups uniquely combined to each other and in fact most of these groups grew out of a very distinct cultural subsections prior to the eighteenth century. For example in germany there was no german state prior to the eighteenth century. It was a conglomeration of many different german and frankish kingdoms that came together to form a modern state, and the modern state is about when these attitudes eventually arose within our society. So it's important to recognize that there's nothing fundamentally human about nationalism, there's nothing that combines these populations in any unique way. Between the fact that they neighbor each other and in some instances share cultural bonds though when you allow for nationalism and when nationalism arises in the way that it has in the last two centuries, it allows for new different cultural bonds to be formed which are frankly exclusive in many ways and most importantly arbitrary in their creation. They're simply made in order to enforce this idea of national identity and national community that doesn't exist and is often a tool of those empowered by nationalism to use that nationalism as a guise for fascism. But firstly, before I get on to that I think it's important to talk about why nationalism is simply a bad political force within the world. Nationalism by its definition is exclusionary. In order to celebrate a nation you must create distinctions between that nation and those around it and while some would argue for a cosmopolitan nationalism that allows for people to celebrate their nation simply because it's something that is diverse and beautiful, such as the united states and the idea of the melting pot, firstly, this isn't how nationalism actually arises in the world. Nationalism is more often in more often the case, nationalism is the force that says: my national identity group, my my ethnicity, my regional nation, any sort of group is is better than other groups that border me, or that there's something that makes them distinct that makes them superior. This false superiority creates a a sense of xenophobia throughout the world, which is one reason why there's, in the, in europe right now there's such a hesitancy to to accept refugees from syria, and from other war torn areas in the middle east and northern and northern africa. This is because there's this idea that there's some sort of benefits that we read from our nation that are exclusive the benefits for our nation. That because we are where we are we have earned the goods and resources that we get from these regions. But we only get these benefits because of the arbitrary nature of where we were born and what our region happens to have and what it can give us. There's no one more deserving of getting these sort of political goods whether it be a stable government or representative democracy than people that are fleeing to these areas as refugees. It's just the luck of where they were born. Given that this is the case we think that nationalism becomes an exclusionary political philosophy that only harms the most disenfranchised people like refugees, who are not able to access the goods that they desperately need. We also think that it creates divisions within a society itself. It means that people that have become part of this communities, say minority groups in in largely white european countries, feel excluded from their own society. Whether it's through ideas of nationalism that simply don't create an image of the nation that includes them, or it's more overt and direct threats. That come from largely far right groups that use nationalism as a guise for fascism. And this the other problem with nationalism. It's that when you create xenophobic senses within a state that creates this sort of false superiority that my nation is better than your nation, it allows for strong man leaders to stand up and say: I'm going to protect the nation. I'm going to ensure the nation rises to its former glory, and these sort of robust senses of pride in the nation allow for these people to get away with crimes and other sorts of corruption that allow them to enrich themselves while at the same time creating strong men groups that create serious threats to democracy not just in developed but also in not just in developing nations but also developed nations such as greece where the xri'si party is rising, and france with marine la pen, in england with braxit and with united states and donald trump. All of these people use nationalism as a way to try and fuel their political anger that their people feel and it only creates more divisions within our society which is frankly contradictory to the global ideas that have been set forth for the past for the past sixty or seventy years of post world war two, peace and prosperity that's occurred. For these reasons we think that nationalism has certainly done more harm than good.
Human Counter Speech	In order to consider, whether nationalism does more harm than good, you must consider the counterfactual: what would have been here had we not had nationalism? We think that, this debate is inherently comparative, in that we think, human beings have an inherent need and desire to group around things that unite them and join them together. This is why in the entire history of mankind, man has always grouped together over certain ideologies, aspects, or whatever it is. Historically, it has taken the form of religion, of monarchism, and of nationalism. Of these alternatives, we think nationalism is by far the best, and we think these alternatives are, in fact, the other options for how life may be. Let's get into rebuttal first. So first, tim says nationalism is exclusionary to other groups, he is correct about that, and then he takes it from that, and says that's why there's xenophobia, that is what he is incorrect about. Xenophobia existed far before nationalism. Religions fought amongst themselves for millennia, so did monarchies who went to war over crown crown and queen, for example. We don't think nationalism caused that. In fact, we think since the rise of nationalism, national wars have gone drastically down. Secondly, he says: minority groups within society feel exclusion, excluded. Again, let's look comparatively. We think a jew, in a christian society, is inherently excluded from that society. We think, an israeli in a american society, can take upon himself aspects of american nationalism, without giving up his religious identity, and thereby allow him to participate in society, more than other groupings would. Lastly, he says: it allows for corrupt leaders. We accept this, it's true. We think it's less so than the alternatives, that are based on a deity. Let's take a look into that. Why is nationalism better? Two reasons: one, based on leaders, second, based on geographic inclusion. First, let's talk about leaders. We think what makes nationalism unique, is that it puts the people in the middle. The comparative of nationalism is various forms of identity, that all include one central leader, be it god, be it chief rabbi, be it a king or a queen. We think that is particularly dangerous, because it allows for that corrupt power, in a significantly more powerful way, than any form of identity based on the nation as a whole. At the point, at which even the leader can be seen to be harming the nation, we think, that nationalism allows groups to protect themselves from corrupt leaders. It is true, that in instances, it also allows them to fall to corrupt leaders, but historically, we think you have far more corrupt leaders under alternative ways of grouping society. So, we think nationalism is better based on the leaders. Let's talk about geographic inclusion. At the point, at which you have two " otherize " some group, because in order to unite yourself with some people, it inherently necessitates creating some form of enemy, and this has been true all throughout history. We think, the best way of doing that, is uniting yourself around the group, based on where you are geographically located. We think that's better, because it's much more difficult to start wars with people who are far away from you. We think that's better, because it's much more difficult to have local tensions, if all of your enemies are far away from you. We think it's better, because all of the reasons, for which humans tend to strive to be in groups, mean that they gain more from these groups, when they are surrounded by these groups. So, nationalism is the best form of grouping together, and grouping together is inherent to human nature. For these reasons, we think nationalism has done far more good than harm.

Table 5: A data sample of the benchmark dataset for the counter speech generation task.

Template for claim detection
Identify whether the given sentence is a claim towards the given topic. Choose from 'claim' or 'non claim'. Sentence: [sentence] Claim: [claim] Label:
Template for stance detection
Identify the stance of the given sentence towards each given target. Choose from 'support', 'attack', or 'neutral' for each target in the target pair. Format the output as a label pair: label1, label2. Sentence: [sentence] Target Pair: [targets] Label Pair:

Table 6: Prompt templates for selected argument mining tasks.

Template for counter argument generation
Identify a premise for a claim and come up with a counter-argument that challenges the validity of that premise. Claim: [claim] Premises: [premises] Counter Argument:
Template for abstractive summarization
Identify the main points and supporting evidence in the document that support the argument being made. Document: [document] Abstractive Summary:

Table 7: Prompt templates for selected argument generation tasks.

Task	Dataset	Train	Dev	Test	Class
Claim Detection	IBM Claims	1500	500	500	2
	IBM Argument	99	9	500	2
	IAM Claims	55044	500	500	2
Evidence Detection	IBM Evidence	3566	500	500	2
	IAM Evidence	56898	500	500	2
Stance Detection	IAM Stance	3371	500	500	2
	IBM Stance	1500	500	500	2
	MTSD	5738	500	500	9
	FEVER	144949	500	500	3
Evidence Classification	IBM Type	291	500	500	3
	AQE Type	7407	500	500	5
Generation	CounterArguGen	0	0	100	-
Summarization	ConcluGen-base	123539	12354	500	-
	ConcluGen-Aspects	122040	12192	500	-
	ConcluGen-Targets	110867	11068	500	-
	ConcluGen-Topic	123538	12354	500	-
	DebateSum-Abstractive	90000	10000	500	-
	DebateSum-Extractive	90000	10000	500	-

Table 8: Dataset statistics.

Model	Claim Detection			Evidence Detection		Stance Detection				Evidence Classification	
	IBM Claims	IBM Argument	IAM Claims	IBM Evidence	IAM Evidence	IAM Stance	IBM Stance	MTSD	FEVER	IBM Type	AQE Type
					Acc.						
Random	50.20	48.00	52.00	49.20	51.00	45.80	47.00	10.00	29.40	33.20	19.20
Flan-T5-XL	74.00	59.00	91.60	69.60	77.40	53.80	27.80	15.40	34.20	73.80	22.20
Flan-T5-XXL	72.00	59.20	88.20	71.60	82.00	61.60	38.00	32.00	34.80	75.20	56.20
Llama-2-7B	32.40	41.20	48.00	45.20	33.60	11.20	7.00	4.20	29.40	6.40	2.40
					F_1						
Random	55.53	52.01	64.28	49.62	58.05	51.76	50.05	12.59	**33.93**	33.51	24.42
Flan-T5-XL	69.07	58.27	90.87	68.69	80.09	43.55	13.46	13.68	27.33	73.66	15.74
Flan-UL2 (20B)	63.49	54.69	89.22	69.17	82.78	60.34	38.75	31.02	28.47	74.83	57.80
Llama-2-7B	24.87	31.34	60.14	42.81	38.50	15.91	7.73	0.68	16.79	10.24	2.59

Table 9: Zero-shot performance of Flan-T5-XL, Flan-T5-XXL and Llama-2-7B on argument mining tasks.

Reference 1	Professional teams shouldn't be required to announce or release the name of their inactive players .
Prediction 1	Teams should not be required to release a list of players that cannot play due to injury or other reasons , as it takes away a strategic advantage for the team.
Reference 2	Free-to-play games are the worst thing happening in the gaming industry today.
Prediction 2	The rise in popularity of free to play games and their associated practices such as micro transactions and pay to win will have a negative impact on the gaming industry as a whole.
Reference 3	I believe that bicyclists need to follow the same set of rules that cars or motorcycles do while on the road, up to and including minimum speed, lane splitting, signaling, and traffic signs. Failing that, they need to stay off of the road.
Prediction 3	Cyclists should be held to the same standards as motorists when it comes to obeying traffic laws and regulations.

Table 10: Examples of the references and predictions from the ConcluGen dataset. Phrases with similar meanings but different expressions are highlighted in pink .

Reference	I'm not sure a wave will necessarily mean the minority party will wrestle away control of the House. I'd argue a wave doesn't just need to be measured by seats won. It can be measured by votes won. It's on this score that Democrats are in a very strong position historically speaking. The problem for Democrats isn't lack of popular support. It's how that support gets translated into seats It's not unusual historically speaking for the minority party to need more than a majority of votes (cast for the two major parties) to win a majority of seats. That's because incumbents tend to outperform the national environment, and the majority party usually has more incumbents running. What is unusual about 2018 is how great the disparity is between votes and seats is likely to be because of geographic sorting and gerrymandering. Usually, the minority party would need to win the House popular vote by a few points to win control, not seven. It just strikes me as unreasonable to expect that Democrats to win a net gain of much more than 23 seats given the vote/seat disparity and that no minority party in a midterm in the modern era has done better than an 8.5-point popular vote win. If Democrats are able to do that, it would be extraordinary.
Longformer	we to I that an electoral will the of and the elections in those I'm not a wave will mean the minority party will wrestle away control of the House. is this to a the and the is a wave? a wave doesn't just need to be by seats won. It can be by votes won. It's on this that Democrats are in a In of over the Democrats minority party) held a 7.5-percentage point on the No minority party has won the national House vote in a election by more than the percentage won it by in the Democrats the national House vote by in The for the election is and the the minority Republicans won it by in and is, Democrats are for a minority party. The for Democrats isn't of It's how that support gets into seats. A 7.5-point win in the national House vote Democrats in the of they need to be for a of seats to win a of seats. The level of disparity between seats and votes won is from it a that about at It's not unusual for the minority party to need more than a of votes for the parties) to win a majority of seats. That's because incumbents to the national and the majority party has more incumbents is about is how the disparity is between votes and seats is likely to be because of and the minority party would need to win the House popular vote by a to win control, not It's this that makes it in to a wave on the of seats won and in terms of control of the it's seats that It's that seats should be included in of a It to that Democrats to win a net of much more than seats the vote/seat and that no minority party in a in the has done than an popular vote win. Democrats are able to do that, it would be
GPT-3.5-Turbo	The article argues that a wave election in the upcoming November elections may not necessarily result in the minority party (Democrats) winning control of the House. While Democrats currently hold a 7.5-percentage point lead on the generic congressional ballot, the level of disparity between votes and seats is likely to be significant due to geographic sorting and gerrymandering. The article suggests that a wave should not just be measured by seats won, but also by votes won. Democrats are doing well for a minority party, but the problem is how that support gets translated into seats. The article concludes that it may be difficult for Democrats to win a net gain of much more than 23 seats given the vote/seat disparity and the historical precedent of no minority party in a midterm election winning the national House vote by more than 8.5 percentage points.

Table 11: Case study 1 of extractive summarization. Incoherent sentences are in red. The effectively extracted important information is in blue.

Reference	Latin American markets are becoming increasingly valuable for Chinese companies because they allow the PRC to expand and diversify its export base at a time when economic growth is slowing. The region has also proven an effective market for Chinese efforts to sell more sophisticated, higher value added products in sectors seen as strategic, such as automobiles, appliances, computers and telecommunication equipment, and aircraft In expanding markets in Latin American the hopes of access to Chinese markets and investments among key groups of business people and government officials in those nations have played a key role in the political will to overcome the resistance China has applied more explicit pressures to induce Latin America to keep its markets open to Chinese goods
Longformer	Access to Latin American Latin American markets are becoming for Chinese companies because they allow the PRC to expand and its export base at a when economic growth is in markets as the United and The region has also proven an effective market for Chinese efforts to sell more higher value added products in sectors as as and and In expanding access for its products through trade with as and and markets in Latin American with as and the PRC has to overcome resistance by and in those In the of access to Chinese markets and investments among of and in those have a in the political to overcome the resistance. it that the Chinese to of the in the could on the and an if an a Chinese Protection of Chinese Investments in and Trade from the China has applied more to induce Latin America to keep its markets open to Chinese goods. It has measures by the Argentine and that it has as protectionist: and, in the of as informal China a causing billion in soy exports and for
`GPT-3.5-Turbo`	Latin American markets are becoming increasingly valuable for Chinese companies as they allow the PRC to expand and diversify its export base. The region has also proven an effective market for Chinese efforts to sell more sophisticated, higher value-added products in strategic sectors. In expanding access for its products through free trade accords with countries such as Chile, Peru, and Costa Rica, and penetrating markets in Latin American countries with existing manufacturing sectors such as Mexico, Brazil, and Argentina, the PRC has often had to overcome resistance by organized and often politically well-connected established interests in those nations. China has applied more explicit pressures to induce Latin America to keep its markets open to Chinese goods, and it has specifically protested measures by the Argentine and Mexican governments that it has seen as protectionist. In the case of Argentina, China began enforcing a longstanding phytosanitary regulation, causing almost $2 billion in lost soy exports and other damages for Argentina.

Table 12: Case study 2 of extractive summarization. Incoherent sentences are in red. The effectively extracted important information is in blue.

Task	Dataset	Setting	k-shot	BERTScore	ROUGE1	ROUGE2	ROUGEL	METEOR
Generation	CounterArguGen	Premises	k=0	77.42	6.40	1.01	4.98	9.81
		Weak Premises	k=0	76.87	5.28	0.62	4.76	8.39
Summarization	ConcluGen	Base	k=0	78.08	3.23	0.98	2.99	8.83
			k=1	$76.75_{0.84}$	$2.49_{0.45}$	$0.57_{0.21}$	$2.27_{0.35}$	$6.35_{1.07}$
			k=3	$76.36_{0.43}$	$1.93_{0.22}$	$0.32_{0.09}$	$1.82_{0.17}$	$5.02_{0.56}$
			k=5	$76.24_{0.27}$	$1.87_{0.22}$	$0.30_{0.09}$	$1.77_{0.19}$	$4.83_{0.46}$
		Aspects	k=0	78.03	4.64	1.41	4.22	10.51
			k=1	$76.55_{1.02}$	$3.55_{0.14}$	$0.81_{0.10}$	$3.22_{0.08}$	$7.86_{0.14}$
			k=3	$76.17_{1.12}$	$2.43_{1.03}$	$0.41_{0.21}$	$2.22_{1.00}$	$5.52_{2.17}$
			k=5	$77.19_{0.34}$	$3.07_{0.69}$	$0.68_{0.27}$	$2.75_{0.63}$	$7.13_{1.24}$
		Targets	k=0	78.10	4.35	1.28	3.98	10.42
			k=1	$77.49_{0.48}$	$3.97_{0.45}$	$0.98_{0.11}$	$3.58_{0.40}$	$9.09_{0.87}$
			k=3	$77.73_{1.25}$	$3.11_{0.45}$	$0.59_{0.22}$	$2.71_{0.37}$	$7.50_{1.04}$
			k=5	$77.41_{1.70}$	$2.91_{0.97}$	$0.41_{0.22}$	$2.54_{0.58}$	$6.75_{1.96}$
		Topic	k=0	77.29	3.78	1.16	3.51	9.13
			k=1	$77.22_{0.58}$	$3.23_{0.33}$	$0.71_{0.13}$	$2.97_{0.26}$	$7.38_{1.11}$
			k=3	$76.75_{0.46}$	$2.20_{0.57}$	$0.34_{0.20}$	$2.02_{0.54}$	$5.41_{1.40}$
			k=5	$76.83_{0.81}$	$2.09_{0.98}$	$0.29_{0.23}$	$1.87_{0.82}$	$5.27_{2.25}$
	DebateSum	Abstractive	k=0	78.55	3.14	0.61	2.71	7.32
			k=1	$77.83_{0.60}$	$2.72_{0.34}$	$0.51_{0.08}$	$2.35_{0.31}$	$6.30_{0.83}$
			k=3	$77.93_{0.49}$	$2.72_{0.33}$	$0.50_{0.09}$	$2.33_{0.31}$	$6.33_{0.77}$
			k=5	$77.89_{0.53}$	$2.73_{0.32}$	$0.52_{0.08}$	$2.36_{0.30}$	$6.37_{0.75}$
		Extractive	k=0	83.71	34.45	24.90	29.74	41.47
			k=1	$84.60_{0.72}$	$36.90_{2.08}$	$27.59_{2.27}$	$31.43_{1.62}$	$44.50_{2.59}$
			k=3	$83.89_{0.84}$	$34.70_{2.24}$	$25.13_{2.96}$	$29.47_{2.28}$	$41.38_{3.97}$
			k=5	$84.91_{0.94}$	$37.99_{2.68}$	$28.96_{3.15}$	$32.73_{2.41}$	$45.97_{3.56}$

Table 13: Performance on argument generation tasks using `Llama-2-7B`.

Task	Dataset	Setting	k-shot	BERTScore	Rouge1	Rouge2	RougeL	METEOR
Generation	CounterArguGen	Premises	k=0	78.01	7.53	0.70	5.59	11.29
		Weak Premises	k=0	78.06	7.87	0.91	5.95	11.85
Summarization	ConcluGen	Base	k=0	78.99	3.95	1.10	3.49	9.73
			k=1	76.05$_{0.43}$	2.79$_{0.18}$	0.55$_{0.10}$	2.42$_{0.15}$	6.55$_{0.65}$
			k=3	76.63$_{0.02}$	2.55$_{0.09}$	0.45$_{0.00}$	2.18$_{0.04}$	6.18$_{0.21}$
			k=5	77.39$_{0.36}$	2.87$_{0.08}$	0.57$_{0.03}$	2.36$_{0.05}$	7.17$_{0.31}$
		Aspects	k=0	78.47	4.04	1.18	3.54	9.68
			k=1	77.38$_{0.40}$	3.90$_{0.23}$	0.83$_{0.08}$	3.38$_{0.22}$	8.57$_{0.22}$
			k=3	77.33$_{0.25}$	3.59$_{0.37}$	0.69$_{0.13}$	3.09$_{0.34}$	7.95$_{0.73}$
			k=5	77.36$_{0.64}$	3.84$_{0.17}$	0.84$_{0.00}$	3.37$_{0.17}$	8.61$_{0.13}$
		Targets	k=0	78.58	4.25	1.27	3.79	10.37
			k=1	78.14$_{0.33}$	4.38$_{0.33}$	1.19$_{0.11}$	3.95$_{0.17}$	10.17$_{0.66}$
			k=3	78.42$_{0.49}$	3.96$_{0.69}$	0.91$_{0.44}$	3.42$_{0.62}$	9.16$_{1.40}$
			k=5	77.69$_{0.34}$	3.10$_{0.68}$	0.64$_{0.24}$	2.71$_{0.67}$	7.54$_{1.10}$
		Topic	k=0	78.71	3.91	1.11	3.48	9.66
			k=1	77.80$_{0.52}$	3.83$_{0.31}$	1.01$_{0.11}$	3.50$_{0.26}$	9.08$_{0.83}$
			k=3	77.92$_{0.77}$	3.18$_{0.40}$	0.59$_{0.08}$	2.72$_{0.28}$	7.62$_{0.73}$
			k=5	77.68$_{0.81}$	2.98$_{0.66}$	0.58$_{0.20}$	2.55$_{0.57}$	7.40$_{1.34}$
	DebateSum	Abstractive	k=0	78.97	3.35	0.62	2.90	7.58
			k=1	78.39$_{0.67}$	2.98$_{0.49}$	0.51$_{0.16}$	2.50$_{0.40}$	6.63$_{1.15}$
			k=3	78.57$_{0.35}$	2.98$_{0.37}$	0.53$_{0.15}$	2.51$_{0.36}$	6.75$_{0.91}$
			k=5	78.38$_{0.95}$	3.09$_{0.22}$	0.55$_{0.10}$	2.59$_{0.22}$	6.71$_{0.96}$
		Extractive	k=0	83.33	32.59	21.45	26.62	37.57
			k=1	84.29$_{1.04}$	35.45$_{3.25}$	25.08$_{3.95}$	29.09$_{3.23}$	41.43$_{4.32}$
			k=3	83.74$_{0.77}$	33.63$_{1.51}$	22.74$_{2.13}$	27.07$_{1.37}$	38.60$_{2.70}$
			k=5	84.65$_{1.15}$	37.03$_{3.57}$	26.81$_{4.58}$	30.72$_{3.32}$	43.16$_{4.55}$

Table 14: Performance on argument generation tasks using `Llama-2-13B`.

Task	Dataset	Setting	k-shot	BERTScore	Rouge1	Rouge2	RougeL	METEOR
Generation	CounterArguGen	Premises	k=0	84.35	10.38	1.03	8.01	5.64
		Weak Premises	k=0	84.39	11.76	1.60	8.71	6.76
Summarization	ConcluGen	Base	k=0	87.37	21.54	8.19	19.03	14.92
			k=1	87.84$_{0.02}$	23.45$_{0.10}$	8.86$_{0.22}$	20.61$_{0.08}$	16.82$_{0.35}$
			k=3	87.92$_{0.05}$	23.94$_{0.27}$	9.06$_{0.50}$	21.00$_{0.36}$	17.22$_{0.64}$
			k=5	87.99$_{0.08}$	24.23$_{0.14}$	9.26$_{0.32}$	21.17$_{0.15}$	17.60$_{0.59}$
		Aspects	k=0	87.41	22.39	8.86	19.81	16.82
			k=1	87.54$_{0.12}$	22.95$_{0.52}$	8.62$_{0.06}$	19.79$_{0.29}$	19.14$_{0.85}$
			k=3	87.73$_{0.09}$	23.83$_{0.19}$	8.97$_{0.26}$	20.44$_{0.25}$	20.72$_{0.41}$
			k=5	87.63$_{0.05}$	23.34$_{0.27}$	8.71$_{0.16}$	20.07$_{0.29}$	19.89$_{0.34}$
		Targets	k=0	87.40	22.58	8.65	19.89	17.06
			k=1	87.62$_{0.09}$	23.46$_{0.39}$	8.90$_{0.25}$	20.33$_{0.31}$	19.69$_{1.26}$
			k=3	87.61$_{0.12}$	23.38$_{0.21}$	8.95$_{0.13}$	20.22$_{0.22}$	19.85$_{0.68}$
			k=5	87.63$_{0.13}$	23.30$_{0.48}$	8.81$_{0.19}$	20.13$_{0.40}$	19.99$_{0.84}$
		Topic	k=0	87.59	22.50	8.32	19.84	16.53
			k=1	87.82$_{0.17}$	23.83$_{0.64}$	8.86$_{0.40}$	20.81$_{0.31}$	18.76$_{1.62}$
			k=3	87.90$_{0.04}$	24.21$_{0.21}$	8.88$_{0.39}$	20.83$_{0.33}$	19.70$_{0.58}$
			k=5	87.92$_{0.09}$	24.02$_{0.24}$	8.89$_{0.54}$	20.85$_{0.44}$	19.52$_{1.03}$
	DebateSum	Abstractive	k=0	85.39	14.97	2.70	12.55	12.73
			k=1	83.81$_{2.75}$	11.42$_{5.76}$	2.06$_{1.09}$	9.55$_{4.78}$	9.76$_{4.74}$
			k=3	85.45$_{0.11}$	14.92$_{0.14}$	2.80$_{0.08}$	12.53$_{0.11}$	12.57$_{0.07}$
			k=5	85.46$_{0.07}$	14.83$_{0.06}$	2.79$_{0.12}$	12.39$_{0.02}$	12.48$_{0.10}$
		Extractive	k=0	85.75	22.28	14.26	18.87	14.60
			k=1	86.15$_{0.35}$	24.36$_{1.77}$	16.34$_{1.77}$	20.89$_{1.70}$	16.26$_{1.44}$
			k=3	86.15$_{0.40}$	24.01$_{2.22}$	16.07$_{2.41}$	20.66$_{2.34}$	15.99$_{2.06}$
			k=5	86.20$_{0.39}$	24.66$_{2.06}$	16.61$_{2.09}$	21.19$_{2.02}$	16.61$_{1.87}$

Table 15: Performance on argument generation tasks using `Flan-UL2`.

Topic	Nationalism does more harm than good
Pipeline 1 (Claim Detection)	Nationalism can actually have positive effects on a country and its citizens. It can promote a sense of unity and pride in one's country, which can lead to increased civic engagement and a desire to work towards the betterment of the nation. Additionally, nationalism can serve as a unifying force in times of crisis or conflict, bringing people together to overcome challenges and protect their country. While extreme forms of nationalism can certainly lead to harm, it is important to recognize that a moderate level of nationalism can have beneficial effects. While it is true that nationalism can be used as a tool for fascism, it is not inherently fascist. National identity and community can be important for promoting unity and a sense of belonging among citizens. Additionally, a strong national identity can help to promote national pride and a sense of shared values and history. It is important to distinguish between healthy nationalism and dangerous nationalism, and to work towards promoting the former while rejecting the latter. While there are certainly negative aspects to nationalism, it is important to recognize that it can also have positive effects. Nationalism can foster a sense of unity and pride among a country's citizens, which can lead to increased social cohesion and a stronger sense of national identity. Additionally, nationalism can be a powerful force for promoting democracy and human rights, as citizens are more likely to demand these things from their government when they feel a strong sense of national identity and pride. Therefore, while nationalism should certainly be approached with caution, it is not necessarily a wholly negative political force. While it is true that nationalism can sometimes lead to exclusionary policies, it is important to note that a sense of national identity and pride can also bring people together and promote unity. Additionally, a strong national economy can provide resources and opportunities for all citizens, including refugees. It is not necessarily nationalism itself that harms disenfranchised people, but rather the actions and policies of those in power who use nationalism as a tool for exclusion and discrimination. Therefore, it is important to address and challenge these harmful actions rather than dismissing nationalism as inherently harmful. While it is true that social media can create divisions within a society, it can also bring people together and foster a sense of community. Social media platforms allow individuals to connect with others who share similar interests and beliefs, regardless of geographical location. This can lead to the formation of online communities that provide support, encouragement, and a sense of belonging. Additionally, social media can be used to raise awareness about important social issues and promote positive change. Therefore, it is important to recognize the potential benefits of social media in promoting social cohesion and unity. While it is true that a strong sense of national pride can sometimes lead to corruption and threats to democracy, it is important to note that not all instances of national pride result in negative consequences. In fact, a healthy sense of national pride can often lead to positive outcomes such as increased civic engagement, a stronger sense of community, and a greater willingness to work towards the common good. Additionally, it is important to recognize that corruption and threats to democracy can arise from a variety of factors, not just national pride. Therefore, it is not fair to solely blame national pride for these issues. Instead, we should focus on addressing the root causes of corruption and threats to democracy, while also promoting a healthy sense of national pride that is grounded in respect for diversity and democratic values. While it is true that nationalism can sometimes lead to divisions within society, it can also serve as a unifying force for a country. Nationalism can create a sense of pride and identity among citizens, which can lead to a stronger sense of community and a shared vision for the future. Additionally, nationalism can be a powerful tool for promoting social and economic progress, as it encourages people to work together towards a common goal. Therefore, while it is important to be mindful of the potential negative effects of nationalism, it is also important to recognize its potential benefits and to use it in a responsible and constructive way. While it is true that nationalism has been associated with negative outcomes such as war and discrimination, it is important to recognize that it has also played a positive role in shaping national identity and promoting unity among citizens. Nationalism has been a driving force behind important social and political movements, such as the fight for independence and civil rights. Additionally, a sense of national pride can inspire individuals to work towards the betterment of their country and its people. Therefore, it is unfair to dismiss nationalism as solely harmful without acknowledging its potential benefits.
End-to-end	While I understand the concerns raised about nationalism, I believe that it is important to recognize the positive aspects of national identity. Nationalism can provide a sense of belonging and community for individuals who share a common culture, language, and history. It can also promote a sense of pride in one's country and encourage individuals to work towards the betterment of their nation. Furthermore, it is important to note that nationalism does not necessarily have to be exclusionary. A cosmopolitan nationalism, as mentioned, can celebrate diversity and promote inclusivity. It is possible to embrace one's national identity while also recognizing and respecting the identities of others. Additionally, the argument that nationalism leads to xenophobia and a reluctance to accept refugees is not necessarily true. While there may be instances where nationalism is used to justify exclusionary policies, it is important to recognize that there are also many individuals and communities who embrace refugees and work towards creating a more inclusive society. Finally, the argument that nationalism leads to the rise of strongman leaders and threats to democracy is not necessarily a result of nationalism itself, but rather a result of individuals who use nationalism as a tool for their own political gain. It is important to hold these individuals accountable for their actions and to work towards promoting a more inclusive and democratic society. In conclusion, while there are certainly concerns and challenges associated with nationalism, it is important to recognize the positive aspects of national identity and to work towards promoting a more inclusive and diverse society.

Table 16: A case study of the data sample (shown in Appendix B) from the benchmark dataset for the counter speech generation task.