LLM DEBATE OPPONENT : Counter-argument Generation focusing on Implicit and Critical Premises

Taisei Ozaki[1], Chihiro Nakagawa[1,2], Naoya Inoue[2,3], Shoichi Naito[4,5]
Kenshi Yamaguchi[5] Atsuhiko Shintani[1]

[1]Osaka Metropolitan University, [2]RIKEN, [3]JAIST, [4]RICOH COMPANY, LTD., [5]Tohoku University
Correspondence: sg23174y@st.omu.ac.jp

Abstract

Debate education is effective in fostering critical thinking skills, an important national issue, but the human cost is a problem. While Large Language Models (LLMs) show promise in automating this process, the optimal approach for targeting critical premises remains unclear. This study proposes methods that specifically focus on implicit and critical premises in counter-argument generation and compares multi-step and one-step implementation approaches. Through evaluation of seven distinct methods using 100 debate topics, we demonstrate that focusing on critical and implicit premises improves counter-argument quality, with one-step methods consistently outperforming multi-step approaches. This superiority stems from better capture of motion spirit, reduced hallucinations, and avoidance of challenging intermediate tasks. Among the methods targeting premises, the Generated and Targeted Premise Attack approach achieved the highest performance in both human expert and automated evaluations. Our findings suggest that counter-argument generation benefits more from integrated approaches that allow LLMs to fully utilize their learned understanding of argumentative patterns. These results provide important insights for developing more effective debate agents and advancing automated argumentation systems.

1 Introduction

In our highly information-oriented society, the development of critical thinking skills[1] is a national priority. It is said that these skills are fostered through debate education. However, debating requires a human cost, such as an opponent and an evaluator. We are therefore developing a debate opponent using Large Language Models (LLM) agents with powerful natural language processing capabilities. It is expected that learners will experience various types of arguments, represented by weakening arguments by denying premises (Sanders, 1974), through this debate against the debate opponents. This exposure to diverse argumentative strategies is expected to enhance the learners' capacity for critical thinking skills (Zhang et al., 2016).

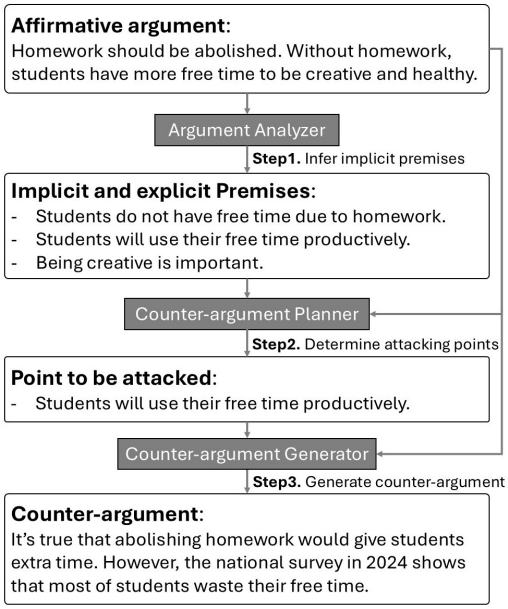

Figure 1: Methods of counter-argument generation

In developing LLMs as debate agents, a critical consideration is their ability to generate counter-arguments. Even in this era of rapidly advancing LLMs, which have seen big progress in text generation capabilities (Lin et al., 2023; Goloviznina et al., 2023; Wang et al., 2023; Chen et al., 2024), research focused on the generation of counter-arguments continues to attract considerable interest within the field. However, the feedback from debate experts[2] suggests that counter-arguments generated by LLMs often lack argumentative strength.

[1]Logical, objective, and unbiased reasoning, characterized by reflective thinking that involves the conscious examination of one's own reasoning processes (Kusumi, 2010).

[2]Members of the Japan Parliamentary Debate Association

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 456–465

In fact, the gpt-3.5-turbo based debate opponent we developed in our preliminary experiments has been defeated by middle school students who have learned to debate in English competitive debate competitions (as of March 2023).

In competitive debate, The strength of the counterargument hinges on the validity of the premise being attacked. For instance, when countering "*Homework should be abolished because it infringes on free time*," challenging the implicit and critical premise that "*free time is inherently more productive*" proves more strong (Walton, 2009).

Therefore, it is important how implicit and critical the premise can be attacked. Several studies focusing on premises in debate exist. Alshomary et al. (2021) proposed a two-step framework using BERT and GPT-2 to directly target and refute key premises, outperforming earlier LSTM-based methods in generating counter-arguments. However, the above proposed method limits the premise to be attacked to explicit ones and does not clarify the criteria or definition of critical premises.

In this study, we proposed a method (see Figure 1) to make LLM imitate the thought process that debate experts implicitly follow when constructing a counter-argument: first, they organize premises that support the affirmative argument, then they decide which premises to attack, and then they create a counter-argument.

In this study, we collaborated with debate experts to independently design a definition of critical premises and proposed a method (see Figure 1) that enables LLMs to mimic the implicit reasoning processes that debate experts naturally employ when constructing counter-arguments. By doing so, we aim to incorporate implicit premises as potential targets for attack, thereby generating counter-arguments with greater argumentative strength.

Our approach consists of three key steps. In the first step, the LLM receives a debate topic along with its corresponding affirmative claim and generates a comprehensive list of premises that support the claim, regardless of whether they are implicit or explicit. In the second step, the model identifies which premises to attack based on the predefined criteria for critical premises. Finally, in the third step, the LLM constructs counter-arguments that specifically target the selected premise.

We evaluated our approach from two key perspectives: (1) whether the target premises for attack should include implicit premises (i.e., whether Step 1 should be performed), and (2) whether providing predefined critical premises impacts performance. As a baseline, we used a simple direct counter-argument generation approach. Furthermore, considering prior research indicating that LLM performance improves when reasoning processes are explicit, as seen in Chain-of-Thought (CoT) prompting (Wei et al., 2023), we investigated whether our method's performance differs when all steps are instructed at one-step versus when each step is executed separately in a multi-step prompting.

Therefore, the purpose of this paper is to evaluate and compare both the multi-step and one-step approaches to counter-argument generation from two perspectives: whether implicit premises are also added to the candidate attack premises or whether the definition of critical premises is used. The contributions of this study are presented below.

- We proposed a method to generate highly strong counter-arguments by having LLM imitate the strategies that human experts use when constructing counter-arguments.
- We showed that even implicit assumptions are candidates for attack assumptions, and that providing critical assumptions is effective in the task of generating counterarguments.
- It directly compares multi-step and one-step generation approaches and provides important insights into the design of LLM-based counter-argument generation systems.

Through comprehensive evaluation involving human experts and automated assessment, we investigate these approaches' effectiveness in generating strong counter-arguments, aiming to contribute to the development of more effective debate agents.

However, our research focuses specifically on the identification and targeting of implicit and critical premises in counter-argument generation, rather than on the procedural approach itself (multi-step or one-step). We suggest that effective counter-arguments should target premises that are critical to the basis of the argument but often left implicit by the arguer. Thus, our key suggestion is to focus on the quality of the premises rather than the generative process.

2 Related Work

LLM-based Counter-Argument Generation. Ozaki et al. (2023) compared GPT-3 counter-arguments with human-crafted ones from Kialo,

showing that LLM responses can match or surpass human outputs in logical coherence. More recent work has leveraged multi-agent interactions among LLMs with distinct personas (Hu et al., 2024) and self-refinement techniques (Madaan et al., 2023; Kao and Yen, 2024; Hu et al., 2023) to further enhance diversity and depth.

Premise-Focused Methods. Attacking premises is a core strategy in debate (Sanders, 1974). Alshomary et al. (2021) proposed a BERT-GPT-2 pipeline for identifying and refuting key premises, outperforming LSTM-based methods. Accounting for *implicit premises* can reveal hidden assumptions, as demonstrated by Boltužić and Šnajder (2016).

Multi-Step Reasoning. Inspired by CoT prompting (Wei et al., 2023) and Zero-shot CoT (Kojima et al., 2023), multi-step methods clarify argumentative structure. Alshomary and Wachsmuth (2023) showed that negating a central claim by selectively attacking premises can improve counter-arguments, though multi-step prompts risk hallucinating premises or misidentifying targets (Ozaki et al., 2024).

Open Challenges. These studies highlight the importance of both explicit and implicit premises, as well as the balance between multi-step and single-step approaches. Our work extends this research by examining how incorporating implicit premises and critical premise definitions, alongside multi-step prompting, affects the strength of LLM-generated counter-arguments.

3 Methods

We categorize our counter-argument generation approaches into multi-step and one-step methods, each reflecting a distinct strategy for producing counter-arguments. The multi-step approach imitates the systematic analytical process of human experts, splitting the generation into phases that can enhance transparency and explainability. By contrast, the one-step approach merges these phases into a single step, while still aligning with the expert-inspired pipeline. As a baseline, we consider a direct counter-argument generation method that does not attempt to replicate expert reasoning. Table 1 compares the main differences. All methods rely on a single LLM agent, use the same system prompt (Table 11), and share generation goals derived from Table 8.

3.1 Multi-step generation

The difference between implicit and explicit assumptions is appended in the Appendix A.

m-Comp: Generated and Targeted Premise Attack Counter-argument Generation

m-Comp comprises three phases. First, it generates a comprehensive list of both implicit and explicit premises underlying the affirmative argument. Second, it selects a single premise to attack by applying the critical premise criteria (Table 9). Finally, it produces a concise counter-argument that focuses on this chosen premise. The entire prompt for this method is shown in Table 12.

m-Targ: Targeted Premise Attack Counter-argument Generation

m-Targ has two phases. Instead of generating premises, it draws on only the explicit premises present in the affirmative argument, chooses one for attack using the critical premise criteria, and then generates a counter-argument focusing on that selected premise. The prompt for this method is in Table 13.

m-Basic: Non-Targeted Premise Attack Counter-argument Generation

m-Basic also proceeds in two phases, similarly selecting a premise from the affirmative argument's explicit statements. However, it does not use critical premise criteria, choosing a premise without that guidance and generating a counter-argument accordingly. The prompt is presented in Table 13.

3.2 One-step Methods

o-Comp, o-Targ, o-Basic

o-Comp, o-Targ, and o-Basic each condense the respective multi-step strategies into one step. o-Comp corresponds to m-Comp, o-Targ to m-Targ, and o-Basic to m-Basic, merging premise consideration and target selection into a single prompt (Table 13). Table 1 summarizes the overall distinctions among these methods.

3.3 Baseline

DirectGen: Direct Counter-argument Generation

OS-0 DG generates a counter-argument in a single step, without explicitly considering any premises. This forms our baseline approach.The prompt is presented in Table 14.

Table 1: Comparison of Methods

Method	Premise Type	Critical Criteria	Steps
m-Comp	Both	input	3
m-Targ	Explicit	input	2
m-Basic	Explicit	no	2
Directgen*	unspecified	no	1
o-Comp	Both	input	1
o-Targ	Explicit	input	1
o-Basic	Explicit	no	1

*baseline

Table 2: Evaluation metrics for Counter-argument

No.	Type	Description
Q1	Ranking	Attacking a more critical premise
Q2	Ranking	Attacking a more implicit premise
Q3	Ranking	The counter-argument is overall stronger
Q4	Choice	Relevance to the topic
Q5	Choice	Logical consistency
Q6	Choice	Multiple supporting reasons
Q7	Choice	Use of specific examples
Q8	Choice	Attacking the affirmative argument's premise

4 Construction of Dataset

We collected debate topics and affirmative arguments from idebate[3], a well-known debate forum. We randomly selected 100 instances from the scraped data and used an LLM (Clade-3.5-sonnet) to refine them into clear, concise sentences while maintaining the original content. Examples are shown in Table 10

5 Experiment

We conducted a comparative evaluation experiment of four counter-argument generation methods. Using a 100-set dataset, we generated counter-arguments using three LLMs: gpt-4o-mini-2024-07-18 (mini") and gpt-4o-2024-05-13 (gpt") from OpenAI[4], and llama-3.1-70b-versatile ("llama") from Meta[5]. We performed automatic evaluation using gpt-4o as evaluator, comparing methods within two groups (multi-step format + baseline and one-step + baseline) using eight evaluation metrics. The metrics were categorized as either choice or ranking type (refer to Table 2), with evaluators reviewing counter-arguments simultaneously within groups. To verify reliability, we conducted parallel experiments with human debate experts, measuring agreement with LLM results. We also directly compared multi-step and one-step approaches through paired evaluations. Calibration was performed using a separate dataset before evaluation experiments.

[3] https://idebate.net/resources/debatabase

[4] https://openai.com/index/openai-api/

[5] https://groq.com/

Table 3: Combined Inter-Rater Agreement Results

Human Experts		
Model	Choice	Ranking
mini	0.53	0.33
gpt	0.34	0.36
llama	0.50	0.30
GPT-4o vs Each Expert		
Model	Choice	Ranking
mini	0.46	0.24
gpt	0.32	0.26
llama	0.43	0.26

Table 4: Probability of ranking in the top of each method evaluated by experts and LLM(40 samples)

	Multi-step			
	m-Comp	m-Targ	m-Basic	Directgen
Q1	0.7583	0.6889	0.5889	**0.8028**
Q2	0.7889	0.6722	0.6528	**0.8806**
Q3	0.7028	0.5806	0.4833	**0.8083**

5.1 Evaluation Metrics

A description of each ranking type evaluation metrics is given below, and a description of the choice type metrics is given in Appendix B.

- **Q1: Attacking a more critical premise** This metric ranks counter-arguments based on how effectively they attack critical premises. Attacks on key, yet under-explained premises are rated higher than those targeting minor or well-defended points.
- **Q2: Attacking a more implicit premise** This metric evaluates how well the counter-argument addresses implicit premises—those assumed but not explicitly stated.
- **Q3:The counter-argument is overall more strong** This metric evaluates the overall effectiveness of the counter-argument, taking into account the importance of the premise attacked, the quality of reasoning, and the overall persuasiveness.

5.2 Inter-Rater Agreement

We calculated the agreement rate of annotations between human expert evaluators (refer to Table 3). Gwet's AC1 was used as the agreement metric(Vach and Gerke, 2023). [6]

When utilizing LLMs as evaluators, the agreement rate with experts decreased by only approximately 0.1 points, indicating that the LLM evaluations did not deviate significantly from those made by human experts.

[6] Krippendorff's α (Krippendorff, 2007) is often used in the NLP field, it was not used in this experiment because it was considered to cause the kappa paradox((Zec et al., 2017)) due to the excessively high agreement rate in the *choice* type indicators.

Table 5: Probability of ranking in the top of each method evaluated by LLM (100 samples)

	Multi-step			
	m-Comp	m-Targ	m-Basic	Directgen
Q1	0.6933	0.4067	0.2033	**0.6967**
Q2	0.6033	0.3567	0.3167	**0.7233**
Q3	0.7300	0.3433	0.1633	**0.7633**
	One-step			
	o-Comp	o-Targ	o-Basic	Directgen
Q1	**0.7000**	0.5367	0.3767	0.5456
Q2	**0.6456**	0.5334	0.5334	0.5454
Q3	**0.6546**	0.5222	0.3567	0.5300

Table 6: Win-rate of one-step against multi-step (100 samples)

Metric	Comp	Targ	Basic
Q1	**0.6078**	**0.6799**	**0.6810**
Q2	**0.7314**	**0.7518**	**0.7849**
Q3	**0.6537**	**0.5612**	0.4946

6 Results and Analysis

The results for ranking-type evaluation metrics are shown in Tables 4, 5[7]. Table 4 shows 40 samples evaluated by experts and GPT-4o; Table 5 shows 100 samples by GPT-4o for multi-step, one-step, and combined methods. We assessed probability of counter-arguments ranking in top positions. Direct comparison results between method pairs in Table 6. Example generation in 17.

In multi-step methods, Directgen achieved highest ranks across Q1-Q3 metrics, followed by m-Comp, m-Targ, m-Basic. In one-step methods, o-Comp ranked highest, Directgen and o-Targ showed equal rates, o-Basic lowest. One-step methods demonstrated superior performance except Q3 comparison between Basic variants.

One-step methods outperform multi-step methods across all metrics. Three key factors contribute to these results. First, better motion spirit capture, as LLMs learn affirmative claims, and counterarguments in proximity within embedding space, while decomposed steps may miss critical premises. Second, reduced hallucination impact, as multi-step processes propagate hallucinations forward ((Zhang et al., 2024),(Nourbakhsh et al., 2022),(Huang et al., 2024)), while one-step generation minimizes impact. Third, premise decision difficulty is a significant challenge. Selecting critical premises has been shown to be difficult even for state-of-the-art LLMs, with (Ozaki et al., 2024) demonstrating that even powerful models

[7]Values averaged across three models. Choice-type metrics in Table 16, Appendix

Table 7: Probability that a premise judged by the LLM to be a valid attack point is also judged by the expert to be a valid attack point (precision score) (Ozaki et al., 2024)

model	Average score
gpt-4	0.79
gpt-3.5-turbo	0.72
llama2-70B-chat	0.59
gemini-pro	0.67
Claude2.1	0.51
Majority baseline	0.62

achieve only about 70% accuracy in selecting effective premises for counter-arguments compared to expert judgments. This research specifically found some disagreement even among human debate experts on what constitutes an optimal target premise, highlighting the inherent complexity of this task. Our observations confirm these findings, with many instances in our experiment showing ineffective premise selection in multi-step approaches.

In a study by Ozaki et al. (2024) that evaluated attack premise selection quality in counter-argument generation, Table 7 shows the precision rates of LLMs compared to expert selections used as the gold-standard. Even the highly capable GPT-4 achieved only approximately 80% accuracy when measured against expert choices, demonstrating the inherent difficulty of the attack premise decision step.

7 Conclusion

This study conducted a comprehensive comparison of different approaches to counter-argument generation using large language models, addressing the challenge of high human costs in debate education while maintaining educational effectiveness. Through evaluation of seven distinct methods across 100 debate topics, we demonstrate that focusing on critical and implicit premises significantly enhances LLMs' ability to generate strong counter-arguments.

Our analysis reveals that one-step methods consistently outperformed multi-step approaches across all evaluation metrics. This superior performance can be primarily attributed to their better capture of motion spirit through LLM's learned associations between topics and counterarguments. Additionally, one-step methods minimize the impact of hallucinations that typically cascade through multi-step processes, while avoiding the challenging task of intermediate premise selection

that often proves difficult even for experienced debaters.

Among the methods targeting premises, o-Comp achieved the highest performance in both human and automated evaluations. Its success stems from the effective consideration of both explicit and implicit premises, combined with clear guidance about critical criteria within a single-step framework. The method's ability to identify and attack core assumptions proved crucial for generating compelling counter-arguments, demonstrating the importance of comprehensive premise analysis in automated argumentation.

These findings contribute significantly to our understanding of how to effectively leverage LLMs in complex argumentation tasks and provide practical insights for developing more effective debate agents. Our results suggest that while decomposed reasoning can be beneficial in many contexts, counter-argument generation benefits more from integrated approaches that allow LLMs to fully utilize their learned understanding of argumentative patterns. These insights pave the way for more accessible and effective debate education systems that can help address the critical need for developing students' critical thinking skills.

8 Limitations and Future Work

Future research should address these limitations through:

- **Dataset Expansion**: Development of various debate data sources beyond idebate, including multi-turn debates and data synthesis by LLM
- **Evaluation Metrics**: Creation of more universal strength rating metrics for counter-arguments that consider argumentative context beyond isolated arguments
- **Hallucination Assessment**: Developing systematic evaluation of factual accuracy in generated counter-arguments, particularly important in debate contexts. As shown by (Ozaki et al., 2024), the premise selection step is especially vulnerable to hallucinations, with LLMs sometimes selecting premises that aren't actually critical to the argument or generating entirely new premises that weren't implied in the original argument. Future work should focus on methods to reduce these hallucinations through knowledge grounding or verification techniques.
- **LLM Analysis**: Comprehensive model-specific effectiveness verification across varying model sizes and architectures
- **Generation Framework**: Multi-turn support and external knowledge incorporation for more practical debate situations
- **Practical Applications**: Integration with debate education platforms and measurement of educational effectiveness through controlled studies

References

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, Online. Association for Computational Linguistics.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. *Preprint*, arXiv:2301.09911.

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany. Association for Computational Linguistics.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2024. Fingen: A dataset for argument generation in finance. *ArXiv*, abs/2405.20708.

Valeriya Goloviznina, Irina Fishcheva, Tatiana Peskisheva, and Evgeny V. Kotelnikov. 2023. Aspect-based argument generation in russian. *COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES"*.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2024. Unlocking varied perspectives: A persona-based multi-agent framework with debate-driven text planning for argument generation. *Preprint*, arXiv:2406.19643.

Zhe Hu, Hou Pong Chan, and Yu Yin. 2023. Americano: Argument generation with discourse-driven decomposition and agent interaction. *ArXiv*, abs/2310.20352.

Qiang Huang, Feng Huang, DeHao Tao, YueTong Zhao, BingKun Wang, and YongFeng Huang. 2024. Coq:an empirical framework for multi-hop question answering empowered by large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11566–11570.

Wei-Yu Kao and An-Zi Yen. 2024. MAGIC: Multi-argument generation with self-refinement for domain

generalization in automatic fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10891–10902, Torino, Italia. ELRA and ICCL.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Klaus Krippendorff. 2007. Computing krippendorff's alpha-reliability. *Departmental papers (ASC)*.

Takashi Kusumi. 2010. *Modern Cognitive Psychology 3: Thinking and Language*. Kitaoji Shobo, Kyoto.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. In *Conference on Empirical Methods in Natural Language Processing*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

Armineh Nourbakhsh, Cathy Jiao, Sameena Shah, and Carolyn Penstein Rosé. 2022. Improving compositional generalization for multi-step quantitative reasoning in question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Taisei Ozaki, chihiro Nakagawa, Naoya Inoue, Shoichi Naito, Takeshi Yamaguchi, Shotaro Amano, and Atsuhiko Shintani. 2024. Premise generation as effective points of counterarguments using large language models. In *Proceedings of the 30th Annual Meeting of The Association for Natural Language Processing*, pages 2681–2686, Japan. The Association for Natural Language Processing.

Taisei Ozaki, Chihiro Nakagawa, Shoichi Naito, Naoya Inoue, Takeshi Yamaguchi, and Atsuhiko Shintani. 2023. Automatic generation of high-quality counterargument papers using large language models. In *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2023, pages 4Xin111–4Xin111.

Gerald H Sanders. 1974. *Debate as a Paradigm for Demonstrating Skills in Argumentation and Logic*. ERIC.

Werner Vach and Oke Gerke. 2023. Gwet's ac1 is not a substitute for cohen's kappa – a comparison of basic properties. *MethodsX*, 10.

Douglas Walton. 2009. Objections, rebuttals and refutations. *Argument Cultures: Proceedings of the 8th OSSA Conference*.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2023. Argument and counter-argument generation: A critical survey. In *International Conference on Applications of Natural Language to Data Bases*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Slavica Zec, Nicola Soriani, Rosanna Irene Comoretto, and Ileana Baldi. 2017. High agreement and high prevalence: The paradox of cohen's kappa. *The Open Nursing Journal*, 11:211 – 218.

Jiawei Zhang, Chejian Xu, Yu Gai, Freddy Lecue, Dawn Song, and Bo Li. 2024. Knowhalu: Hallucination detection via multi-form knowledge based factual checking. *ArXiv*, abs/2404.02935.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Appendix

A Definition of keywords

Table 8: Definition of keywords

Debate: A structured discussion on a specific topic, where participants are divided into the affirmative and negative sides. The affirmative side argues for the benefits that can be gained by accepting the topic, while the negative side emphasizes the potential drawbacks.
Counter-argument: Taking the opposing stance to the argument, critically identifying weaknesses, inaccuracies, and a lack of supporting evidence in the reasoning of the argument, without creation of a new argument from scratch.
Premise: All the implicit or explicit conditions and propositions that the subject of a argument assumes in order to establish the validity of that argument. Explicit: each sentence that constitutes the argument. Implicit: Unstated premises necessary for the argument to hold

For example: In an argument about the abolition of homework, When the affirmative side argues that "*Homework should be abolished because it takes away students' free time. The long hours of forced study at school, extended to after-school hours, inhibits the students' free time to develop their own ideas. This may indirectly prevent future innovation.*", An explicit premise is each statement that "*Homework should be abolished because it takes away students' free time.*", "*The long hours of forced study at school, extended to after-school hours, inhibits the students' free time to develop their own ideas.*", "*This may indirectly prevent future innovation.*". On the other hand, Implicit assumptions include the following examples, "*Free time is important and valuable in student development*","*Time to develop original ideas leads to future innovation*".

Table 9: Definition of Critical premises

Foundational Importance: It should be foundational to the affirmative argument, supporting a key aspect of their arguments. Attacking the root of the opponent's argument is generally more critical.
Moderate Vulnerability: It should be moderately poorly explained or insufficiently supported in the affirmative argument, as the underlying premises of the opponent's argument are generally better explained and may be preemptively refuted.

For example: In an argument about social media regulation, a foundational premise might be "social media causes significant harm to mental health." This premise is both crucial to the argument (Foundational Importance) and often lacks comprehensive evidence (Moderate Vulnerability). Attacking the above premise and negating a premise that supports elements close to the root of the opponent's argument can significantly weaken their stance. Conversely, a premise that is underexplained in the opponent's argument is easier to attack from various perspectives. Generally, premises that support the core elements of an affirmative argument are well-explained, while those further from the core are often less thoroughly explained. Therefore, the ideal premise for rebuttal should be somewhat close to the core and not fully explained - a middle ground.

B Choice evaluation metrics

Our evaluation framework employs five different choice-type evaluation metrics, each designed as a binary classification task in which the counter-argument under evaluation meets or does not meet the metrics.

Q4: Relevance to the topic This metric evaluates whether the counter-argument stays focused on the debate topic. Effective counter-arguments must directly engage with the main issue, avoiding digressions into unrelated matters.
Q5: Logical consistency This metric evaluates the logical flow of the counter-argument. A strong counter-argument should progress naturally, with no unreasonable leaps or inconsistencies in reasoning.
Q6: Multiple supporting reasons This metric evaluates whether the counter-argument presents multiple reasons to strengthen its claim. Providing several well-reasoned points typically enhances the persuasiveness of the argument.
Q7: Use of specific examples This metric evaluates the use of concrete examples to support the counter-argument. Specific, relevant examples make the argument more tangible and convincing.
Q8: Attacking the premise on which affirmative argument stands This metric evaluates whether the counter-argument directly attacks a key premise that the affirmative argument depends on. A strong counter-argument must challenge a critical foundation of the opponent's reasoning.

C Example of Topic and Affirmative argument

Table 10: sample of dataset

Topic:Should male infant circumcision be considered a form of child abuse?
Affirmative argument: Performing surgery on infants without medical necessity is inherently risky and irresponsible. The Royal Dutch Medical Association has stated that no medical organization worldwide can definitively prove a medical need for infant circumcision. They emphasize that due to the lack of medical necessity and the genuine risk of complications, extremely stringent requirements should be in place for providing information and advice on this procedure. Despite this, circumcision is routinely performed globally, often by individuals with minimal medical training, and is frequently accepted by parents based on religious beliefs rather than medical evidence. This practice exposes infants to unnecessary surgical risks without clear medical benefits, which can be considered a form of child abuse.

D Prompts of each methods

Table 11: System prompt

system prompt: *You are a skilled debater.Your final objective is to make a high-quality counter-argument against an affirmative argument provided on a specific topic. To achieve this: You are not required to create a new argument from scratch. Take the opposite stance of the affirmative argument. To make an counter-argument means to carefully point out the weaknesses, inaccuracies, and lack of evidence in the reasoning of the claim. You may also be asked to complete several other tasks along the way. Consider these tasks as necessary steps to achieve the final objective.*

Table 12: m-Comp prompt

Premise generation step: *topic:#topic# affirmative argument:#argument# Thoroughly analyze the given affirmative argument on the given topic. Identify and list all premises supporting the affirmative argument, with a special emphasis on:1.Explicit premises: Clearly stated premises or sentences.2.Implicit premises: Unstated premises necessary for the argument to hold. Please output only the listed premises.*
Premise decision step: *Select the most suitable premise to attack for your counter-argument from the list of premises. The ideal premise should meet the following criteria: 1. Foundational Importance: It should be foundational to the affirmative argument, supporting a key aspect of their arguments. Attacking the root of the opponent's argument is generally more critical. 2. Moderate Vulnerability: It should be moderately poorly explained or insufficiently supported in the affirmative argument, as the underlying premises of the opponent's argument are generally better explained and may be preemptively refuted. Please output only the premise you chose.*
Counter-argument generation step: *Please make a concise and brief counter-argument to the affirmative argument, that attacks the specific premise you chose. Please output only the text of your counter-argument.*

Table 13: m-Targ and m-Basic prompt

m-Targ prompt
Premise decision step: *topic:#topic# affirmative argument:#argument# premise list:#premise list# Select the most suitable premise to attack for your counter-argument from the premise list. The ideal premise should meet the following criteria: 1.Foundational Importance: It should be foundational to the affirmative argument, supporting a key aspect of their arguments. Attacking the root of the opponent's argument is generally more critical. 2.Moderate Vulnerability: It should be moderately poorly explained or insufficiently supported in the affirmative argument, as the underlying premises of the opponent's argument are generally better explained and may be preemptively refuted.Please output only the premise you chose.*
Counter-argument generation step: *Please make a concise and brief counter-argument to the affirmative argument, that attacks the specific premise you chose. Please output only the text of your counter-argument.*

m-Basic prompt
Premise decision step: *topic:#topic# affirmative argument:#argument# premise list:#premise list# Select the most suitable premise to attack for your counter-argument from the premise list. Please output only the premise you chose.*
Counter-argument generation step: *Please make a concise and brief counter-argument to the affirmative argument, that attacks the specific premise you chose. Please output only the text of your counter-argument.*

Table 14: Directgen prompt(baseline)

Directgen prompt
Counter-argument generation step:*topic:#topic# affirmative argument:#argument# Please make a concise and brief counter-argument to the affirmative argument. Please output only the text of your counter-argument.*

Table 15: o-Comp prompt

topic: #topic# *affirmative argument: #argument#* *First Thoroughly analyze the given affirmative argument on the specified topic. Identify all premises supporting the affirmative argument, including:* *Explicit premises: Clearly stated assumptions or claims.* *Implicit premises: Unstated assumptions necessary for the argument to hold.* *Next choose the most suitable premise to attack for your counter-argument from the premises. The ideal premise should meet the following criteria:* *Foundational Importance: It should be foundational to the affirmative argument, supporting a key aspect of their arguments. Attacking the root of the opponent's argument is generally more critical.* *Moderate Vulnerability: It should be moderately poorly explained or insufficiently supported in the affirmative argument, as the underlying premises of the opponent's argument are generally better explained and may be preemptively refuted.* *Finally, please provide a concise, straightforward counter-argument to the affirmative argument, attacking the specific premise you chose.* *Please output only the text of your counter-argument.*

o-Targ prompt
Counter-argument generation step:*topic:#topic# affirmative argument:#argument# premise list:#premise list# First, select the most suitable premise to attack for your counter-argument from the premise list. The ideal premise should meet the following criteria: 1.Foundational Importance: It should be foundational to the affirmative argument, supporting a key aspect of their arguments. Attacking the root of the opponent's argument is generally more critical. 2.Moderate Vulnerability: It should be moderately poorly explained or insufficiently supported in the affirmative argument, as the underlying premises of the opponent's argument are generally better explained and may be preemptively refuted.Next, Please make a concise and brief counter-argument to the affirmative argument, that attacks the specific premise you chose. Please output only the text of your counter-argument.*

o-Basic prompt
Counter-argument generation step:*topic:#topic# affirmative argument:#argument# premise list:#premise list# First, Select the most suitable premise to attack for your counter-argument from the premise list. Next, Please make a concise and brief counter-argument to the affirmative argument, that attacks the specific premise you chose. Please output only the text of your counter-argument.*

E Results of choice evaluation

The results of the choice evaluation performed by a human expert are shown below. Each value is the percentage of each metrics satisfied.

Table 16: Results of choice evaluation

Relevance to the topic (40 sample)				
metric	m-Comp	m-Targ	m-Basic	Directgen
Q4	0.97	0.97	0.96	0.96
Q5	0.96	0.94	0.86	0.95
Q6	0.74	0.54	0.47	0.87
Q7	0.29	0.28	0.24	0.24
Q8	0.89	0.90	0.90	0.83

F Examples of multi-step and one-step counter-argument generation

Table 17: Examples of multi-step and one-step counter-argument generation

Topic
Feminism is necessary or not
Affirmative argument
In the United Kingdom, one in four people experiences domestic violence (DV). Globally, reports of rape are increasing, yet conviction rates are declining. In the United States, women earn only 77 cents for every dollar earned by men. When filtered for racial minorities, the disparity becomes even more pronounced. These issues highlight the deep-rooted gender problems in society. Feminism still has many objectives to achieve, including women's rights, safety, and equal economic opportunities.
m-Comp)
The claim that women in the United States earn only 77 cents for every dollar earned by men is a misleading statistic that does not take into account various factors such as occupation, education, experience, and working hours. Using this statistic as a fundamental argument for the necessity of feminism oversimplifies the complex nature of wage determination and misrepresents the issue.
Directgen
The issues that are claimed to remain under feminism's purview are, in fact, already being addressed. Overemphasizing these problems can obscure the need for a more inclusive approach that addresses the rights of all genders, including men and non-binary individuals.