Dynamic Knowledge Integration for Evidence-Driven Counter-Argument Generation with Large Language Models

Anonymous ACL submission

Abstract

This paper investigates the role of dynamic external knowledge integration in improving counter-argument generation using Large Language Models (LLMs). While LLMs have shown promise in argumentative tasks, their tendency to generate lengthy, potentially unfactual responses highlights the need for more controlled and evidence-based approaches. We introduce a new manually curated dataset of argument and counter-argument pairs specifically designed to balance argumentative complexity with evaluative feasibility. We also propose a new LLM-as-a-Judge evaluation methodology that shows a stronger correlation with human judgments compared to traditional reference-based metrics. Our experimental results demonstrate that integrating dynamic external knowledge from the web significantly improves the quality of generated counter-arguments, particularly in terms of relatedness, persuasiveness, and factuality. The findings suggest that combining LLMs with real-time external knowledge retrieval offers a promising direction for developing more effective and reliable counter-argumentation systems. Data and code publicly available.[1]

1 Introduction

Argumentation in Natural Language Processing (NLP) is becoming an increasingly active area of research, driven by the natural human tendency to express disagreement with claims or viewpoints expressed by individuals during information exchanges. In fact, it is becoming an indispensable tool in many application domains such as public policy, law, medicine, and education (Stab and Gurevych, 2017; Eger et al., 2018; García-Ferrero et al., 2024; Yeginbergen et al., 2024; Sviridova et al., 2024).

[1] https://anonymous.4open.science/r/counter-argument-generation/

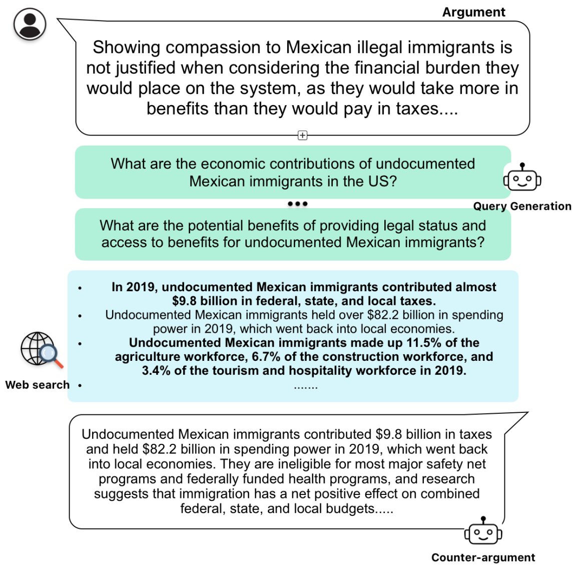


Figure 1: Our approach to counter-argument generation integrating dynamic external knowledge.

It is possible to distinguish two main research lines in Argumentation in NLP: (i) argument mining, which involves analyzing unstructured texts to automatically identify and extract argumentative elements (Cabrio and Villata, 2018; Stab and Gurevych, 2017; Yeginbergen et al., 2024); (ii) argument generation, which focuses on generating argumentative texts using external knowledge sources (Hua et al., 2019; Schiller et al., 2021) or summarizing key argumentative points (Roush and Balaji, 2020; Syed et al., 2021; Chen et al., 2024).

This paper investigates whether dynamic integration of external knowledge helps Large Language Models (LLMs) to improve counter-argument generation. Counter-argument generation seeks to develop an effective framework for presenting alternative perspectives to an argument while ensuring the correctness of the message and incorporating factual evidence (Wachsmuth et al., 2017, 2018). LLMs have shown promising potential to deal with debates or any disagreements by solely relying on the parametric knowledge encoded in the model (Chen et al., 2024; Alshomary and Wachsmuth, 2023). Moreover, most of the LLMs have security safeguarding to avoid harmful interactions by any means (Bai et al., 2022; Ouyang et al., 2022; Touvron et al., 2023). While maintaining the harmlessness is recommendable, it is also important to be reasonable, persuasive, and grounded with a factual basis when arguing (Khan et al., 2024).

Previous work on counter-argument generation has focused on specific aspects of the argument (Schiller et al., 2021; Saha and Srihari, 2023), integrating different external news sources (Hua et al., 2019), or by attacking different evidences of the initial argument (Jo et al., 2020; Chen et al., 2024). However, no previous work has considered integrating dynamic knowledge to inform counter-argument generation. Furthermore, we argue that existing datasets for counter-argument generation consist of too long (Hua and Wang, 2018; Hua et al., 2019; Chen et al., 2024) or too short (Schiller et al., 2021; Lin et al., 2023) counter-arguments which makes it difficult to accurately evaluate their quality. Thus, although modern LLMs tend to generate rather lengthy essay-style responses that may be highly persuasive but lack coherence, logic, and factual evidence, restricting them to short single sentences is insufficient to study the complexity and desired pragmatic characteristics of a good counter-argument. Finally, the manual evaluation of counter-argument quality is a challenging, time-consuming, and highly subjective task that requires knowledge of the subject matter and credible evidence to support or refute it (Wachsmuth et al., 2017; Wang et al., 2017; Hua et al., 2019). In this sense, we believe that traditionally used (Alshomary et al., 2021; Chen et al., 2024) reference-based automatic metrics such as BLEU, METEOR or BERTScore fail to accurately capture the nuanced relationship between the generated counter-argument and the human judgment.

In order to address these issues, we propose using retrieved external knowledge from dynamic sources without any limits to particular outlets or databases, along with LLMs, to generate efficient, concise evidence-based counter-arguments. Furthermore, we present a new curated dataset of concise and structured human-generated argument and counter-argument pairs in which the length of the counter-arguments is enough to study the main argumentative aspects while facilitating manual and automatic evaluation. While this paper employs both manual and automated evaluation methods, we introduce a novel automated evaluation approach using LLMs-as-a-Judge, designed to optimize correlation with human assessments.

Using the new dataset and dynamically retrieved external evidence, this work aims to answer the following research questions (RQ): **RQ1**: Does the integration of dynamic external knowledge into LLMs help to generate better counter-argumentation? **RQ2:** Which automatic evaluation method correlates better with human judgments? **RQ3:** To what extent do LLMs use retrieved external evidence in producing counter-arguments?

Figure 1 illustrates our proposed framework. First, we automatically generate queries (in the form of critical questions) that challenge the main points of the argument or claim and feed those queries to the web search. Next, related evidence is retrieved, and, lastly, the claim and the retrieved evidence are provided as context to the LLM to generate a counter-argument.

The main contributions of our paper are the following: (i) we publicly share a new dataset of manually curated arguments and counter-arguments; (ii) we introduce a new method to dynamically integrate external knowledge retrieved from the web in LLM-based counter-argument generation; (iii) experimental results show that the generation of counter-arguments with LLMs is improved through the integration of dynamic external knowledge, with factual evidence demonstrating a particularly significant impact on pragmatic aspects including relatedness, persuasiveness, and factuality and (iv), our automatic evaluation based on LLM-as-a-Judge reveals a higher correlation with human judgments compared with reference-based metrics such as BLEU, METEOR, or BERTscore.

2 Related Work

Prior research in efficient and persuasive argument generation has approached the problem from various perspectives focusing on different aspects of arguments. For instance, Jo et al. (2020) investigated the ability of machine learning approaches to detect attackable sentences in the arguments, and they concluded that automatic approaches are more stable in this task than human annotators depending on the subjectivity, topic, and tone of the argument. Alshomary et al. (2021) focused on generating counter-arguments by pinpointing and challenging weak premises. Schiller et al. (2021) produce arguments by specifying the desired aspect and stance in a sentence-level setting via Conditional Transformer Language Model (CTRL) (Keskar et al., 2019). Similarly, Saha and Srihari (2023) proposed a method to control both the topic and stance of an argument while enriching it with factual evidence using an encoder-decoder language model. Alshomary and Wachsmuth (2023) propose a counter-argument generation based on the stance of the conclusion of the argument. Lin et al. (2023) employed large language models for sentence-level counter-argument generation, implementing a Chain-of-Thought methodology. Finally, Chen et al. (2024) evaluated LLMs in several argument mining and generation tasks showing the potentiality of LLMs for this particular task.

The use of external sources has demonstrated its effectiveness in the generation of alternative perspectives. Wachsmuth et al. (2018) analyzed the question of retrieving the best counter-arguments when no prior knowledge is persistent. Hua and Wang (2018) introduced a framework that incorporated information retrieval, but their approach was limited to using the Wikipedia database as the external source. However, Wikipedia primarily contains static factual information, which may not align with the dynamic nature of arguments. To address this limitation, Stab et al. (2018) expanded the scope by indexing all documents from the Common Crawl database for argument retrieval. Hua et al. (2019) proposed an enhanced framework that leverages databases from news outlets alongside Wikipedia to retrieve evidence and improve the quality of the generated counter-arguments.

All these efforts mainly refer to static databases as external sources, meaning that all the documents containing the evidence of the argument in question should be parsed in advance. Moreover, previous argument generation with external knowledge was proposed before the LLMs entered the race. In our work, we believe that we should test the capabilities of LLMs and that, in our ever-changing dynamic world, it is not efficient to rely on a pre-defined set of documents as an external source for factual and persuasive counter-argument generation. Instead, we propose to integrate knowledge from the whole internet as a source for finding factual evidence to generate better counter-argumentation.

3 Data

In order to perform our experiments and avoid input and output length inequality in the data, we constructed a new corpus for the evaluation of counter-argument generation. Previous work often focused on either (too short) sentence-level (Lin et al., 2023; Schiller et al., 2021) or (too long) paragraph-level (Hua et al., 2019; Hua and Wang, 2018) generation.

Given the capabilities of modern LLMs, it was essential to define clear input and output data to ensure accurate, robust, and fair comparisons during evaluation. Thus, our objective is to create a dataset of argument/counter-argument pairs that would meet specific criteria. We argue that (i) generating single-sentence claim-based counter-arguments is insufficient to accurately assess the quality of counter-arguments produced by LLMs and (ii), too long counter-arguments make it extremely difficult to properly evaluate the argumentative quality of the generated text. Our analysis revealed that Large Language Models (LLMs), when not given explicit length constraints, tend to generate verbose, essay-like responses that frequently deviate from true argumentative form, lack substantiating evidence, and demonstrate poor coherence with the original argument while overemphasizing persuasive elements. To address these limitations, we propose generating counter-arguments with a maximum length of three sentences, focusing on conciseness, factual content, and direct alignment with the input argument.

With this aim in mind, we constructed a new dataset of argument and counter-argument pairs using the `CANDELA` corpus as a basis (Hua et al., 2019). The corpus consists of debates and discussions on various controversial topics from the `r/ChangeMyView`[2] subreddit. The corpus is centered around real-world online debates where users post their opinions and evidence for any controversial topic and expect other users to provide reasoning for an alternative perspective.

`CANDELA` is available in a format split by sentenced, tokenized, and lowercase, making the reconstruction of the corpus necessary. To address this, we employed LLMs to convert the data back into a coherent, human-readable format.

Once the data was fully reconstructed in a human-readable format, we summarized all arguments and counter-arguments. To avoid any bias we choose to use a language model different from

[2] https://www.reddit.com/r/changemyview

	# sentence	# words
arguments		
Original	16	1996
Intermediate	3	511
Final	3	387
counter-arguments		
Original	30	4478
Intermediate	5	845
Final	3	460

Table 1: Average number of sentences and words in arguments and counter-arguments in the original, summarized (intermediate), and final versions of the data.

those used in our experimentation, namely, Llama-3.1-70B instruct (Dubey et al., 2024). All summaries were then manually verified against the original data and re-summarized where needed to ensure the semantic and pragmatic correctness of the content.

While the corpus includes data from real-world interactions and reflects arguments from natural exchanges of information, we found that not all topics were suitable for counter-argument generation with external knowledge. Specifically, topics lacking a scientific or factual foundation, such as philosophical questions or deeply subjective topics, often trigger LLMs to produce generic, safe responses due to their safety guardrails.

To mitigate potential quality issues, we implemented a manual filtering process for the corpus, retaining only those arguments that demonstrated both high quality and direct relevance to the subject matter, specifically selecting instances where the incorporation of factual information would meaningfully contribute to argument generation.

Finally, we further manually refined the summaries to follow a structured argumentative format, emphasizing components such as the main claim, supporting evidence, and examples where applicable. This process resulted in a dataset of 150 high-quality 3-sentence paragraph-level argument and counter-argument pairs. The final version of the corpus used in the experiments can be seen in the Appendix 3. The average distribution of sentences and tokens in the data is shown in Table 1.

4 Experimental setup

In this section, we describe the methodology for generating counter-arguments with dynamic knowledge integration for evidence-driven counter-argument generation. Preliminary experiments revealed that relying on static pre-defined document sets (such as Wikipedia) for factual evidence retrieval often yields incomplete information regarding specific argument topics, resulting in incoherent and unreliable evidence. We determined that external sources must contain topic-specific content, opinions, and observations directly relevant to the events described in the argument, particularly since claims may reference recent events that post-date the last update of pre-parsed sources. Consequently, our experimental design incorporates *dynamic* web-based data as the primary information source to address these limitations.

External knowledge is obtained through the web search provided by the Cohere API[3]. The process involved automatically generating five queries (averaging 87 words each) designed to challenge the validity and veracity of the original argument's claims and premises by questioning key points that required factual substantiation. These queries were sequentially submitted to the web search engine, and the retrieved results, averaging 5,496 words in length, were incorporated as contextual information in the final prompt presented to the language model. The model then generated counter-arguments based on both the original argument and the retrieved contextual information.

To assess the role of external knowledge in counter-argument generation with LLMs, we performed a comparative analysis using two system configurations: one incorporating external information and another relying solely on the model's parametric knowledge. The model in the latter configuration received only the original argument as input and was tasked with generating counter-arguments using its internal knowledge base exclusively. This experimental design enabled us to evaluate whether LLMs display better performance when provided with external evidence for counter-argument generation, as measured by comparing the quality of outputs between the two configurations.

This experimentation is performed with two instruct LLMs with strong results on text generation tasks in two different configurations: (i) the LLM using only the claim and its parametric knowledge as prompt and (ii) the LLM with dynamic external knowledge. This results in the following four models:

[3] https://cohere.com/

- **Command R+**: Command-R+, a 104B parameter model from Cohere For AI (2024).
- **Command R+ with external knowledge**: Command-R+ 104B with external evidence retrieved using Cohere's API for web search.
- **Mistral-7B-Instruct**: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023).
- **Mistral-7B-Instruct with external knowledge**: Mistral-7B-Instruct-v0.3 with external evidence retrieved via Cohere's API for web search.

Importantly, all experiments are conducted in a zero-shot inferential setting to assess the real capabilities of LLMs in counter-argument generation. This setup ensures a fair and robust evaluation of their performance in generating meaningful, well-reasoned, and factual counter-arguments.

4.1 Evaluation

Following Hua et al. (2019), we assess the quality of the generated counter-arguments using a pointwise 3-point Likert scale across five key dimensions: Opposition, Relatedness, Specificity, Factuality, and Persuasiveness.

It should be noted that the evaluation of generated counter-arguments is inherently subjective. While human evaluation is the gold standard, it is time-intensive, costly, and prone to individual biases (Wachsmuth et al., 2017; Hua et al., 2019; Chen et al., 2024). To address these limitations, along with human evaluation, we will also provide two types of automatic evaluation. First, using reference-based metrics such as BLEU, METEOR or BERTscore previously used in counter-argumentation generation (Alshomary et al., 2021; Chen et al., 2024). Second, we propose to use the LLM-as-a-Judge approach. Leveraging the consistency, scalability, and efficiency of the LLMs, this method enables rapid and reproducible scoring across the five dimensions. To the best of our knowledge, we are the first to propose an LLM-based evaluation for counter-argument generation.

Taking the gold standard counter-argument dataset built in Section 3, we generate a counter-argument for each claim using the four models listed above. The five counter-arguments (gold reference included) are then evaluated by human annotators and LLM-as-a-Judge annotators. More specifically, they are asked to assess each dimension by categorizing the example as *unsatisfactory (score: 1)*, *moderately satisfactory (score: 2)*, or *highly satisfactory (score: 3)* for each of the five dimensions listed above. Finally, reference-based metrics were computed by comparing the generated counter-arguments with the gold reference.

We will also compute the correlation between both automatic evaluation methods and human judgments, as establishing strong alignment is crucial for ensuring valid comparisons and reducing reliance on human evaluators, ultimately leading to a more robust and practical evaluation framework.

Human evaluation. We recruited human evaluators through the crowdsourcing platform Prolific[4] to assess the quality of a sample of 75 generated counter-arguments across the five predefined dimensions. We performed two rounds of evaluations of three participants to ensure that the obtained manual judgments were of high quality. We set test questions in the questionnaire to determine whether the evaluations were performed fairly. Subsequently, we selected the four participants who correctly passed the test. All evaluators were compensated above the minimum rate recommended by the platform. An example of instructions provided for manual evaluation can be found in Appendix C.

LLM-as-a-Judge. We employ two state-of-the-art open LLM-as-a-Judge models, namely, Prometheus (Kim et al., 2024) and JudgeLM (Zhu et al., 2023), and one proprietary model, Claude 3.5 Sonnet[5]. The models are prompted with the same set of instructions for evaluation as those used for human annotators.

Reference-based evaluation. Following the previous evaluations on counter-argument generation (Hua et al., 2019; Lin et al., 2023; Chen et al., 2024) we evaluate using reference-based overlap and similarity metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2019). BLEU and ROUGE compare the overlap between the counter-argument and the claim whereas METEOR considers also synonyms, paraphrases, and word stems. Finally, BERTScore takes into account the semantic context and meaning of the text, going beyond surface-level word matching.

Ranking. Based on the evaluation scores per dimension, we obtain rankings of each the 5 candidate counter-arguments (4 automatically generated and the gold reference) by summing the scores for

[4] https://www.prolific.com/

[5] https://www.anthropic.com/claude/sonnet

each counter-argument, as provided by the human and LLM-as-a-Judge evaluators. Formally, let the evaluation process involve n counter-arguments, m dimensions, and e evaluators. Let $Score_{i,j,d}$ represent the score assigned by evaluator j for counter-argument i on dimension d.

The total score for counter-argument i, summed across all dimensions for evaluator j, is given by:

$$T_{i,j,d} = \sum_{d} Score_{i,j,d}$$

The final ranking for counter-argument i, combining scores across all m dimensions for each evaluator j, is calculated as:

$$R_{i,j} = Rank(\sum_{d=1}^{m} T_{i,j,d})$$

Where:

- $T_{i,j,d}$ is the total aggregated score for counter-argument i for dimension d by evaluator j.
- m is the number of dimensions, 5 in our case.
- $R_{i,j}$ is the rank of the counter-argument i calculated by summing $T_{i,j,d}$ from evaluator j.

The counter-arguments are ranked in ascending order according to their calculated $O_{i,j}$ values, where lower scores correspond to superior performance. This ordinal ranking methodology effectively normalizes individual scoring variations and minimizes evaluator bias, as it accounts for potential systematic differences in scoring tendencies among evaluators who might consistently assign either higher or lower scores, thereby ensuring a fair comparison.

The evaluation using reference-based metrics involves calculating scores for counter-arguments generated by each LLM in comparison to the gold reference, which enables the establishment of relative rankings among the LLMs' outputs.

5 Results

We first calculate the correlation of the automatic metrics with human judgments. We then used the best automatic metrics to compute the overall rankings for the gold counter-arguments and those generated by the four models. Finally, we analyzed dimension-wise rankings obtained in the manual evaluation.

Correlation with Human Judgments. We would like to stress that manual counter-argument evaluation is a highly subjective and tedious process (Wachsmuth et al., 2017; Hua et al., 2019; Chen et al., 2024). Thus, ensuring that our automatic evaluation method correlate with human judgments is crucial. Therefore, we computed the correlation between every evaluation metric on the evaluator sample set of 75 examples, including human evaluation, using the method described in Section 4. The Spearman's rank correlation coefficients (ρ) are reported by Figure 2. First, it can be observed that the three LLM-as-a-Judge methods obtain the highest correlation with human judgments (row marked in red). Thus, while JudgeLM and Prometheus obtain a strong correlation, Claude 3.5 Sonnet is the best method with a ρ of 0.82 (very strong correlation).

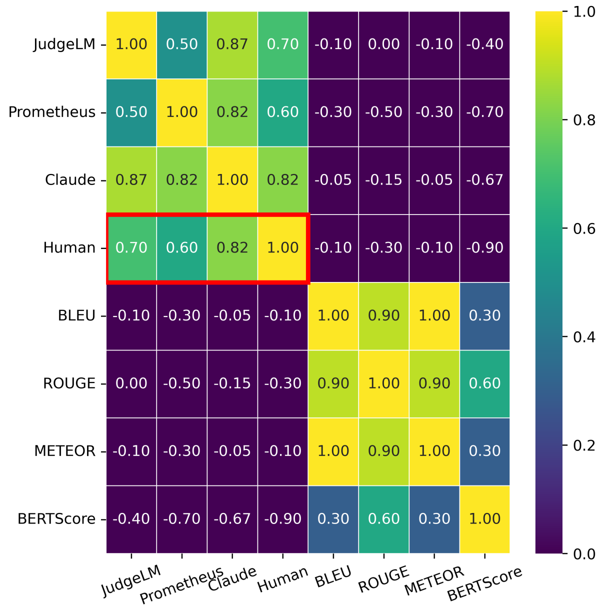


Figure 2: Heatmap showing Spearman's rank correlation coefficients between human evaluation and automatic metrics, including LLM-as-a-Judge metrics. The row marked in red represents the correlation of all the evaluation metrics to human preference.

Reference-based metrics show a poor correlation with both human judgments and LLM-as-a-Judge evaluation methods. This suggests that reference-based metrics might be inadequate for evaluating counter-argument generation quality, as they fail to capture the essential dimensions established for human evaluation and notably disregard the context of the original argument or claim being countered.

Among the LLM-as-a-Judge methods, the lowest correlation was observed between JudgeLM

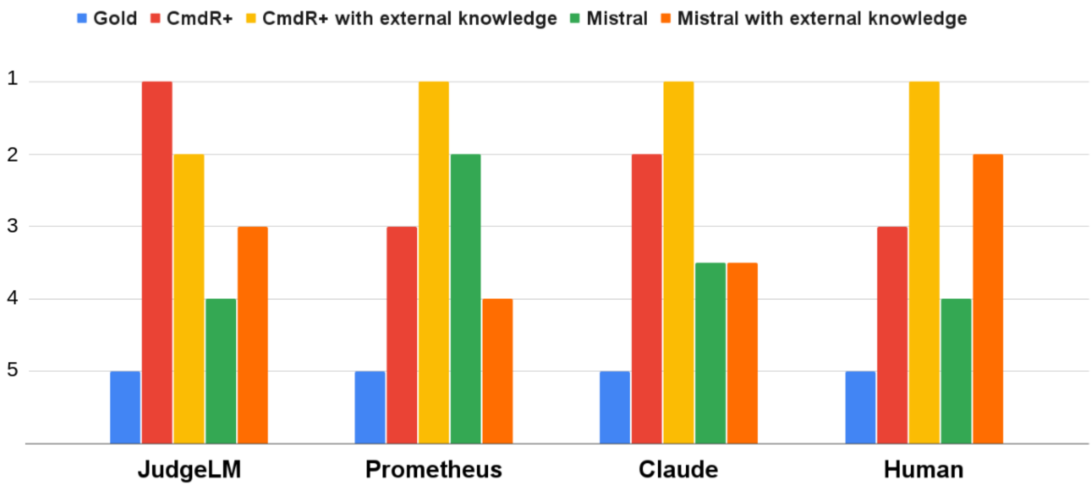


Figure 3: Human and LLM-as-a-Judge evaluation results.

and Prometheus. During a more detailed analysis, we found that these two judges had the highest disagreement in evaluating Opposition and Persuasiveness. Prometheus tends to be stricter on Opposition while JudgeLM does the same on Persuasiveness. Thus, each model assigned lower scores than the other for those particular dimensions.

Overall rank evaluation. Figure 3 reports the overall rankings obtained by summing all the scores from each dimension that each evaluator provided (as described in Section 4). The first observation is that 3 out of 4 LLM-as-a-Judge agree that Command R+ with external knowledge generates the best counter-arguments. Moreover, according to the manual evaluation, both Command R+ and Mistral-7B-Instruct with external knowledge are ranked at the top. Interestingly, results with respect to the ranking of other models vary, but all four evaluators agree that the human-generated counter-arguments (gold reference) are ranked worst.

Table 2 shows the scores obtained by the reference-based metrics. While Command R+ with external knowledge remains the best model, it is easy to see that the rankings are not aligned with respect to human preferences. Furthermore, and although not directly comparable, the obtained scores are in the range of previous work evaluating counter-argument generation with these metrics (Chen et al., 2024).

Taking into account the high correlation of LLM-as-a-Judge methods with human preferences, we also evaluated the four models over the full corpus of 150 arguments and counter-arguments pairs. The results from each LLM-as-a-Judge evaluator can be found in the Appendix B (Figures 5, 6, 7, 8).

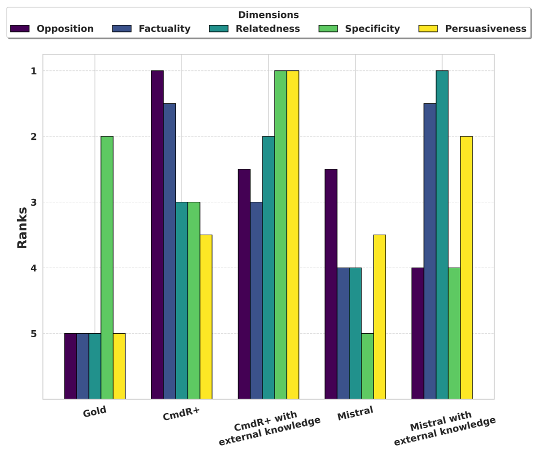


Figure 4: Per dimension ranking from manual evaluation.

Point-wise evaluation. Figure 4 illustrates the average dimension-wise rankings from manual evaluation. We can see that Command R+ excels in Opposition compared to others, whereas both Command R+ and Mistral-7B-Instruct with external knowledge are valued the best on Persuasiveness and Relatedness. Moreover, Mistral-7B-Instruct with external knowledge and Command R+ obtain the highest Factuality scores. Overall, we can conclude that, when external evidence is provided to the LLMs, the overall quality of the counter-argument improves. Nevertheless, the performance of generating counter-arguments based on parametric knowledge is quite high, especially for the larger models (Command R+).

Model	BLEU	ROUGE	METEOR	BERTScore	Avg
Command R+	20.35	18.36	16.12	**86.38**	35.30
Command R+ with external knowledge	**20.80**	**18.67**	**16.81**	86.15	**35.60**
Mistral-7B-Instruct	17.36	15.93	13.96	86.23	33.37
Mistral-7B-Instruct with external knowledge	17.30	16.58	14.36	86.29	33.63

Table 2: Results with reference-based metrics; best overall model per metric in **bold**; best model per family underlined.

6 Analysis of Results

To answer RQ3, we checked whether the generated counter-arguments indeed used the provided external knowledge, or whether the retrieved evidence is already incorporated in the parametric knowledge of LLMs.

We used the sentence transformer model 'gte-base-en-v1.5' (Li et al., 2023; Zhang et al., 2024) to calculate sentence-level cosine similarity between the generated outputs and provided external evidence. Our methodology involved segmenting both the external information and generated counter-arguments into individual sentences for pairwise comparison, followed by ranking the similarity scores in descending order. Based on manual verification, we established a threshold of 70% similarity as indicative of successful external knowledge integration. Our analysis revealed that Command R+ with external knowledge effectively utilized external evidence in 82% of its generated counter-arguments, while Mistral-7B-Instruct with external knowledge demonstrated such integration in 51% of the cases.

Our manual analysis of similarity scores revealed that, when similarity scores exceed 70%, models directly integrate provided external knowledge, while scores between 65-70% result in partial incorporation mixed with additional information. Notably, this partial incorporation behavior predominantly occurs with sensitive topics such as personal reputation, religion, ethics, economics, or politics, where models tend to generalize rather than use specific factual evidence, leading to lower Opposition dimension scores. Command R+ with external knowledge frequently exhibited this behavior, though interestingly, the same model without external knowledge produced more generalized responses that evaluators often found more plausible than fact-based counterarguments.

The topics of high controversy trigger LLMs to start generation by acknowledging the stance of the input argument which exhausts the space for including the retrieved evidence, even when explicitly prompted not to do so. Moreover, such counter-arguments were evaluated better than the output that starts the alternative perspective straight from the factual examples.

7 Conclusion

In this study, we examined the effect of incorporating externally dynamically retrieved evidence into LLMs in counter-argument generation using web search as an external knowledge source. Our results on a newly created gold dataset show that, while LLMs with external knowledge improved their counter-argument generation, their reliance on it varies, and in some cases, they generate responses with parametric knowledge which obtained better scores.

We propose LLM-as-a-Judge to automatically evaluate the quality of counter-arguments with better correlation scores with respect to human judgments than previously used reference-based metrics.

Through qualitative analysis, we found that model behaviour shifts when dealing with sensitive or controversial topics. In these cases, LLMs tend to provide more generalized responses rather than directly integrating factual evidence. Interestingly, such responses were often rated more favourably, suggesting a preference for plausibility and coherence over strict factual accuracy.

Our findings highlight the complexities of integrating external knowledge into LLM outputs. While retrieval-augmented generation (RAG) can enhance factual consistency, models may still prioritize linguistic fluency and alignment with social norms. Future work should focus on refining strategies to ensure that external knowledge is utilized effectively, particularly in contexts that require precise and evidence-based argumentation.

8 Limitations

Our study has some limitations. We focused on two LLMs, both with and without integrated external knowledge, to compute the agreement between human and LLM judgments. Including more models would have significantly strengthened our conclusions. Equally, including other languages should allow for more generalizable results. However, as far as we know, there are no counter-argument generation datasets for languages beyond English. Therefore, one of the short-term objectives of the NLP research community should be to address this glaring gap.

Additionally, due to the effort required for manual assessment, we evaluated only a sample of the generated counter-arguments rather than the entire dataset. Future work could explore more scalable evaluation methods to extend the analysis.

Finally, the LLM-generated counter-arguments may be affected by potential data contamination, where topics and examples of the arguments used in our experiments may overlap with the training data of the LLMs we used. Investigating task contamination is far from trivial but it should be included in any future work.

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.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based Counter-Argument Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *Preprint*, arXiv:2204.05862.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *IJCAI*, volume 18, pages 5427–5433.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the Potential of Large Language Models in Computational Argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330.

Cohere For AI. 2024. c4ai-command-r-plus-08-2024.

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

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An Open-Source Multilingual Text-to-Text LLM for the Medical Domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177.

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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, and Chris Reed. 2020. Detecting Attackable Sentences in Arguments. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–23, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with More Persuasive LLMs Leads to More Truthful Answers. In *Forty-first International Conference on Machine Learning*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. *arXiv preprint arXiv:2405.01535*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. *Preprint*, arXiv:2308.03281.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720, Singapore. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7.

Sougata Saha and Rohini Srihari. 2023. ArgU: A Controllable Factual Argument Generator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8373–8388, Toronto, Canada. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-Controlled Neural Argument Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2024. CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18463–18475.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating Informative Conclusions for Argumentative Texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2024. Argument Mining in Data Scarce Settings: Cross-lingual Transfer and Few-shot Techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11687–11699.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. *Preprint*, arXiv:2407.19669.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. *arXiv preprint arXiv:2310.17631*.

A Example of data

Table 3 shows the sequential methodology employed in generating the corpus. *Original* presents an argument from the publicly available CANDELA corpus that originally is in lowercase and segmented in sentences and tokens. After its reconstruction to a human-readable format, in the *Intermediate* step, the argument is summarized by means of the language model Llama-3.1-70B instruct. The row *Final* presents the version after manual elaboration which is made available in this paper.

Original	["we", "should", "n't", "worry", "about", "being", "compassionate", "to", "mexican", "illegal", "immigrants", "the", "same", "way", "we", "do", "n't", "worry", "about", "being", "uncompassionate", "to", "the", "rest", "of", "the", "world", "'s", "poor", ".", "i", "am", "specifically", "referring", "to", "poor", "immigrants", "who", ",", "based", "on", "current", "tax", "codes", ",", "will", "take", "far", "more", "in", "benefits", "than", "they", "would", "pay", "in", "taxes", ".", "it", "has", "nothing", "to", "do", "with", "skin", "color", "...", "if", "you", "have", "millions", "of", "white", "people", "suddenly", "all", "working", "manual", "labor", "jobs", "and", "below", "you", "now", "have", "a", "lot", "of", "people", "not", "paying", "many", "taxes", "into", "the", "system", "and", "being", "eligible", "to", "take", "a", "lot", "out", ".", "why", "do", "people", "argue", "we", "need", "to", "be", "''", "compassionate", "''", "when", "with", "that", "same", "logic", "you", "could", "argue", "we", "are", "n't", "being", "compassionate", "for", "not", "all", "living", "a", "minimalist", "life", "and", "sending", "all", "our", "wealth", "to", "africa", "until", "there", "are", "no", "more", "starving", "people", "?", "what", "makes", "Mexico", "so", "deserving", "of", "our", "aid", "but", "not", "other", "countries", "?", "logically", "i", "'m", "sure", "the", "people", "clamoring", "to", "he", "compassionate", "and", "let", "all", "the", "poor", "immigrants", "in", "(", "i.e.", "making", "the", "immigration", "process", "easier", "or", "amnesty", ")", "realize", "we", "could", "n't", "support", "the", "entire", "world", ",", "so", "why", "is", "mexico", "special", "?", "what", "do", "you", "consider", "our", "breaking", "point", "for", "percentage", "of", "us", "poor", "(", "i", "think", "we", "'re", "already", "at", "it", ")", "to", "where", "there", "'s", "more", "money", "going", "out", "to", "social", "programs", "and", "tax", "breaks", "than", "there", "is", "coming", "in", "?", "at", "what", "point", "does", "the", "super", "pro", "immigration", "person", "decide", "the", "us", "is", "''", "full", "''", "?"]
Intermediate	The writer argues that showing compassion to Mexican illegal immigrants is not justified when considering the financial burden they would place on the system, as they would take more in benefits than they would pay in taxes. The sustainability of immigration policies is based on economic impact rather than emotions. Uncertainty about the threshold at which the U.S. would be considered "full" and unable to support more immigrants
Final (Ours)	Showing compassion to Mexican illegal immigrants is not justified when considering the financial burden they would place on the system, as they would take more in benefits than they would pay in taxes. This argument is not based on skin color, but rather on the economic impact of a large influx of low-income workers.

Table 3: Example of the data. *Original* is the original publicly available data. *Intermediate* is the summary generated by Llama-3.1-70B. *Final* is the final version after manual refinement and the data used in the experiments.

B Per dimension evaluation ranks

In this appendix we provide evaluation ranks per opposition, factuality, relatedness, specificity, and persuasiveness dimensions for the following LLM-as-a-Judge models: JudgeLM in Figure 5, Prometheus in Figure 6 and Claude in Figure 7. The overall ranks over the whole dataset by LLM-as-a-Judge are shown in Figure 8.

Results, especially those computed using Claude 3.5 Sonnet, align with those obtained in Section 5 with a human-annotated sample of the data. Thus, 8 shows that Command R+ with external knowledge is also ranked first, gold reference, and Mistral-7B-Instruct the worst while the ranks for Command R+ and Mistral-7B-Instruct with external knowledge vary according to the judge.

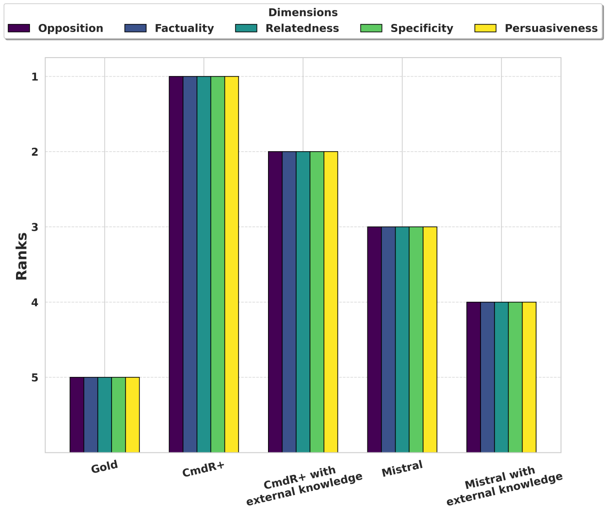


Figure 5: JudgeLM per dimension ranks.

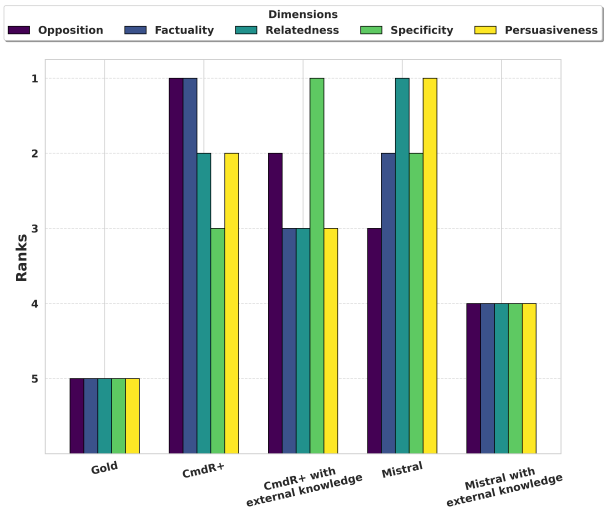


Figure 6: Prometheus per dimension ranks.

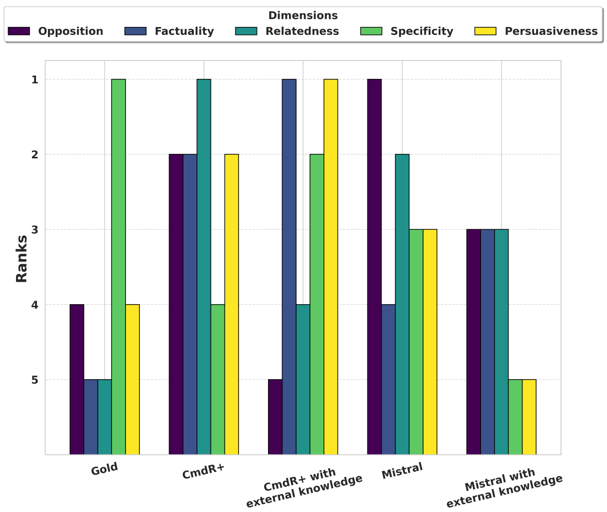


Figure 7: Claude per dimension ranks.

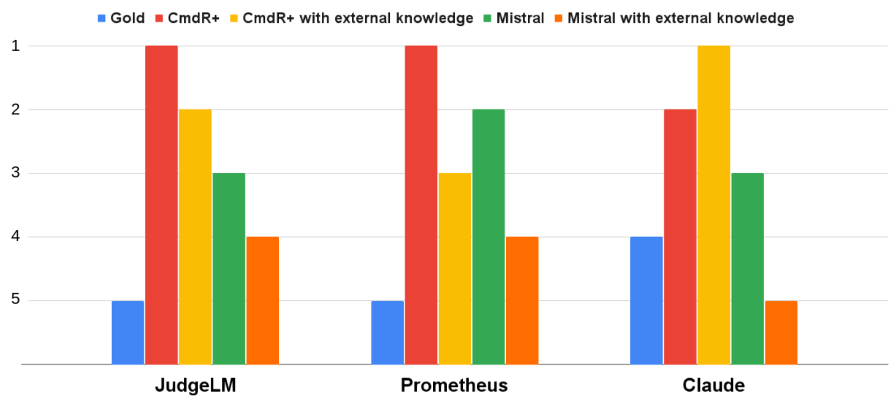
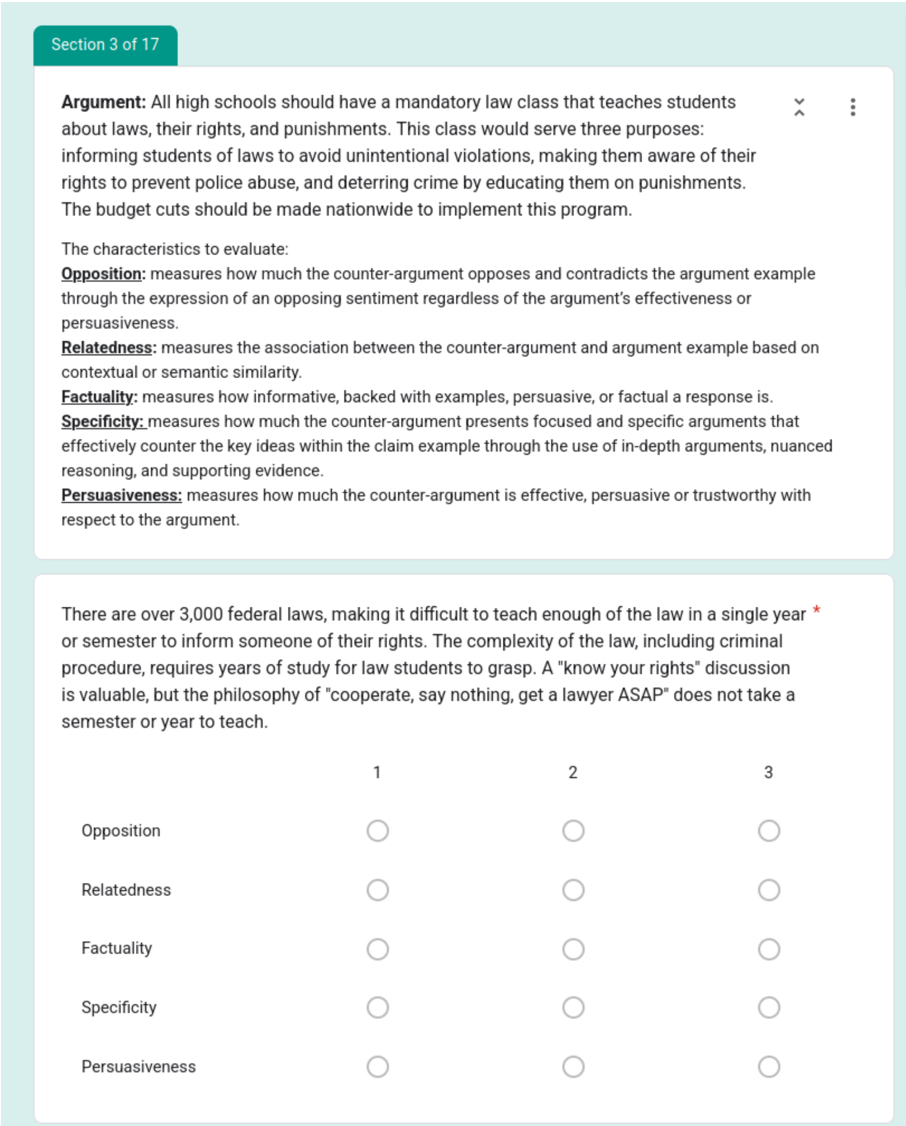


Figure 8: LLM-as-a-Judge evaluation results over the full dataset.

C Instruction example for human evaluation.

An example of the instructions provided to the human evaluators in the Prolific platform is shown in Figure 9. Firstly, the argument is presented accompanied by a description of the five dimensions (opposition, relatedness, factuality, specificity, persuasiveness) that are evaluated in the paper. Secondly, each of the counter-arguments is presented and the evaluator has to select a value between 1 for *unsatisfactory*, 2 for *moderately satisfactory* and 3 for *highly satisfactory* in each of the dimensions.



Section 3 of 17

Argument: All high schools should have a mandatory law class that teaches students about laws, their rights, and punishments. This class would serve three purposes: informing students of laws to avoid unintentional violations, making them aware of their rights to prevent police abuse, and deterring crime by educating them on punishments. The budget cuts should be made nationwide to implement this program.

The characteristics to evaluate:
Opposition: measures how much the counter-argument opposes and contradicts the argument example through the expression of an opposing sentiment regardless of the argument's effectiveness or persuasiveness.
Relatedness: measures the association between the counter-argument and argument example based on contextual or semantic similarity.
Factuality: measures how informative, backed with examples, persuasive, or factual a response is.
Specificity: measures how much the counter-argument presents focused and specific arguments that effectively counter the key ideas within the claim example through the use of in-depth arguments, nuanced reasoning, and supporting evidence.
Persuasiveness: measures how much the counter-argument is effective, persuasive or trustworthy with respect to the argument.

There are over 3,000 federal laws, making it difficult to teach enough of the law in a single year or semester to inform someone of their rights. The complexity of the law, including criminal procedure, requires years of study for law students to grasp. A "know your rights" discussion is valuable, but the philosophy of "cooperate, say nothing, get a lawyer ASAP" does not take a semester or year to teach. *

	1	2	3
Opposition	○	○	○
Relatedness	○	○	○
Factuality	○	○	○
Specificity	○	○	○
Persuasiveness	○	○	○

Figure 9: An example of instructions provided to the human evaluators.