Contextualizing Argument Quality Assessment with Relevant Knowledge

Anonymous ACL submission

Abstract

Automatic assessment of the quality of arguments has been recognized as a challenging task with significant implications for misinformation and targeted speech. While real-world arguments are tightly anchored in context, existing computational methods analyze their quality in isolation, which affects their accuracy and generalizability. We propose *SPARK*: a novel method for scoring argument quality based on contextualization via relevant knowledge. We devise four augmentations that leverage large language models to provide feedback, infer hidden assumptions, supply a similar-quality argument, or give a counter-argument. *SPARK* uses a dual-encoder Transformer architecture to enable the original argument and its augmentation to be considered jointly. Our experiments in both in-domain and zero-shot setups show that *SPARK* consistently outperforms existing techniques across multiple metrics.

1 Introduction

Reliable analysis of arguments in natural language holds the promise to support applications such as automated grading (Ludwig et al., 2021), and tackling misinformation and targeted speech (Alhindi, Tariq, 2023). Computational argument analysis has been relatively popular through tasks like argument extraction (Chakrabarty et al., 2019), evidence mining (Rinott et al., 2015), relation assignment (Trautmann et al., 2020), writing support (Stab and Gurevych, 2014b) and claim generation (Bilu and Slonim, 2016). A particularly challenging task is argument quality assessment (Fromm et al., 2022), which addresses the cogency, effectiveness, and reasonableness of an argument (Wachsmuth et al., 2017b) pertaining to a topic. Assessing the quality of the argument involves analyzing the objective evidence, relevant assumptions, and structural soundness, making the overall task difficult.

Research on argument quality assessment has focused on extracting textual patterns using various learning frameworks and content features (Lauscher et al., 2022). It has been widely recognized that contextualizing arguments with implicit knowledge, such as extracting claim revisions (Skitalinskaya et al., 2021) and generating explicit conclusions (Gurcke et al., 2021) can be informative for reasoning models. However, we note that: 1) these methods fail to generalize to novel arguments where this information is not available, and 2) no prior work has considered jointly a comprehensive set of such contextualization strategies.

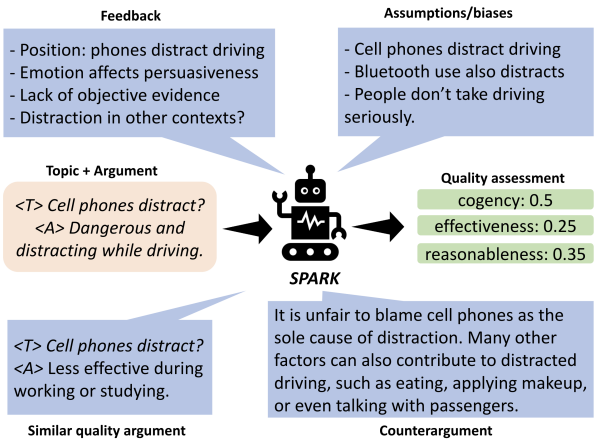

Figure 1: Overview of *SPARK*.

To bridge these gaps, we propose a novel framework called *SPARK* (**S**coring the **P**ragmatics of **A**rguments via **R**elevant **K**nowledge), which incorporates augmentation strategies based on a large language model (LLM), GPT 3.5 (OpenAI, 2022), and elements from argumentation literature (Nickerson, 2020; Mulyati et al., 2023; Harvey, 2009), specifically, feedback, assumptions, arguments with similar quality, and counter-arguments. *SPARK* processes the original argument and topic and its augmentations separately using a dual-encoder Transformer architecture with a multi-head cross-attention layer. We demonstrate the effectiveness of *SPARK*'s augmentations and architecture using both in-domain and out-of-domain datasets. Our entire code is made available at *(anonymized)*.

2 Background

Task formulation. Inspired by prior work (Gretz et al., 2020; Lauscher et al., 2020), we formalize argumentation quality assessment as a regression task of predicting the quality of a natural language argument. Given a topic and an argument, we consider three quality indicators (Lauscher et al., 2020): 1) *cogency*, which evaluates the relevance and sufficiency of the argument's premise in relation to the conclusion, 2) *effectiveness*, which measures the argument's persuasive power based on factors like arrangement, clarity, and appropriateness, and 3) *reasonableness*, which determines the argument's ability to resolve the debate's issue (Wachsmuth et al., 2017a). The overall quality of an argument can be estimated by averaging these three metrics (Gretz et al., 2020).

Connection to prior studies. The introduction of benchmarks for argument quality (Stab and Gurevych, 2014a; Gretz et al., 2020) has inspired various methods based on logistic regression (Ghosh et al., 2016), fully connected and recurrent neural networks (Habernal and Gurevych, 2016), and fine-tuned Transformers (Toledo et al., 2019). Hulpus et al. (2019) explained that contextualizing an argument with implicit knowledge is essential to understanding its quality. Lauscher et al. (2022) categorized knowledge used by current argument assessment research so far under linguistic, task-specific, and argument-specific sections. To mimic human reasoning over arguments, prior work has incorporated users' prior beliefs as predictors of argument persuasiveness (Durmus and Cardie, 2018), trained classifiers for different audience groups (El Baff et al., 2020), utilized user history to predict persuasion (Al Khatib et al., 2020), augmented arguments with supporting or refuting documents (Marro et al., 2022), and augmented arguments with visual features (Hasan et al., 2021). Most similar to *SPARK*, Skitalinskaya et al. (2021) leverage comparison between revisions of the same claims, while Gurcke et al. (2021) generate conclusions to assess argument sufficiency. However, revisions are rarely provided for novel arguments, whereas generated conclusions are argument-specific and may not generalize well (Gurcke et al., 2021). Addressing prior work limitations, *SPARK* implements four well-motivated augmentation strategies to enhance novel arguments and utilizes an attention-based dual encoder model for effective reasoning.

3 *SPARK*

Augmentation strategies. We devise four augmentation techniques to contextualize arguments. We generate the augmentations by prompting GPT-3.5 (OpenAI, 2022) (see appendix for details).

Feedback. Constructive feedback in the form of comments and suggestions helps comprehension and domain knowledge acquisition (Mulyati et al., 2023). We hypothesize that assessing argument strengths and weaknesses helps argument ranking, as in Figure 1, where feedback identifies emotional appeal, insufficient evidence, and generalization. We prompt the LLM to generate writing feedback for a topic-argument pair in a zero-shot setting.

Assumptions. Unstated assumptions frequently introduce bias in arguments (Nickerson, 2020). Making assumptions explicit can reveal these hidden biases, which may aid in assessing argument persuasiveness and relevance. One such assumption in Figure 1 is that people do not take driving seriously. We employ an LLM to extract the argument's underlying assumptions in a zero-shot setting.

Similar-quality instance. Inspired by prior work on claim revisions (Skitalinskaya et al., 2021), we hypothesize that retrieving arguments with similar quality at training time leads to generalizable model learning. For this purpose, we derive a synthetic argument with similar reasonableness, cogency, and effectiveness to the original one (Figure 1). We generate this synthetic argument in a few-shot setting, where the LLM has access to example arguments alongside their quality scores covering the full 1-5 range. Since this augmentation uses ground-truth information that is not available during inference, we randomly replace synthetic arguments with *None* at training time with a probability of $P = 0.5$, thus familiarizing the model with the absence of similar arguments during testing where it only sees *None*. This technique is similar to distillation (Hinton et al., 2015), where the encoder's (student) goal is to utilize the LLM's (teacher) ranking-based argument generations (soft labels) to learn the argument quality ranking task.

Counter-arguments. Counter-arguments provide objections, alternatives, and doubts of a skeptical reader (Harvey, 2009). We expect that contrasting the strengths and weaknesses of two opposite arguments will aid quality assessment. The example counter-argument in Figure 1 makes a firm claim that alternative activities (e.g., eating) lead to distractions, and solely blaming phones is unfair. We

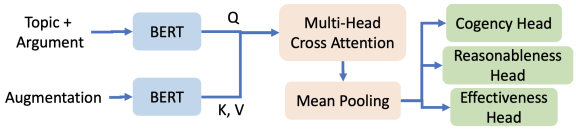

Figure 2: Dual BERT encoder architecture.

ask the LLM to provide a counter-argument for a topic and an argument in a zero-shot setting.

Dual-encoder architecture. To consider the argument together with the augmentations, we employ a dual BERT encoder (Figure 2) as an improvement to the architecture by Gillick et al. (2018). The first encoder embeds the topic and argument, whereas the second embeds the augmentations. The second encoder can store individual augmentations or their concatenation, arbitrarily fixed to *Similar quality argument [SEP] Feedback [SEP] Assumptions [SEP] Counter-argument*. Notably, the dual encoder can effectively store all of the augmentation data without truncating information in practically all cases (see subsection A.4 for the augmentation lengths). We use a multi-head cross-attention layer (Vaswani et al., 2017) to enable the model to weigh each augmentation according to the argument-topic pair. We pass the attention outputs to a mean pooler, whose output is fed into three separate regressor heads, one per quality metric.

4 Experiments

4.1 Baselines

Scoring models. We compare our dual BERT with a standard BERT model to study the effect of disjoint embeddings. We use bert-base-uncased (Devlin et al., 2019) as an encoder. We compare dual BERT to XLNet (Yang et al., 2019), as they can both handle more than 512 tokens. Finally, we utilize GPT-3.5 in a zero-shot setting to gauge the model's ability to accomplish the task directly, without a dual encoder. We provide GPT-3.5 with the definitions of each metric and prompt it to individually rate each argument by a float between 1 and 5 with respect to the topic.

Alternative augmentation strategies. We evaluate the impact of using all augmentations together or one at a time, against ablated baselines without augmentations. We include two alternative augmentation methods: Wikipedia paragraphs extracted using dense passage retrieval (DPR) (Karpukhin et al., 2020), and augmentations generated using smaller models, namely, Flan-T5-XL (Chung et al., 2022) and Llama-2 (7B) (Touvron et al., 2023).

4.2 Datasets and Evaluation

We use *GAQCorpus* (Lauscher et al., 2020) as our training dataset for its diversity of domains (reviews, QA, and debates) and quality metrics (cogency, effectiveness, and reasonableness). We also use *IBM-30K* (Gretz et al., 2020) to test *SPARK*'s generalization on out-of-domain data. For IBM-30K, we perform weighted averaging (WA) of the three metric scores, which is supported by the high correlation between IBM-30K's WA and the GAQ-Corpus metrics (Lauscher et al., 2020). We report the Pearson (ρ) and Spearman (σ) correlation coefficients between the predictions and ground truth.

4.3 Results

We investigate *whether augmentations help language models assess the quality of arguments more effectively* (Q1); *how augmentation strategies compare to each other* (Q2); and *whether human quality judgments align with their model utility* (Q3).

Effect of augmentation on in-domain performance (Q1). Table 1 shows that the overall best-performing combination uses Dual BERT with all four augmentations combined, which improves the Spearman correlation over the baseline BERT by 0.08-0.17 across the three metrics. The improvement is the largest for effectiveness, where the Spearman correlation increases by 61%. While both single BERT and XLNet benefit from *SPARK*'s augmentations as well, their performance is consistently lower than using the dual encoder. GPT-3 alone often performs better than the other baselines but lags significantly behind *SPARK*, showing the importance of the dual encoder. Among the augmentations, the benefit of our four augmentations declines when using a smaller generative model (Flan T5 and Llama-2), augmenting via DPR, or using the dual BERT with a masked second encoder. Thus, merely adding text to the second encoder does not by itself bring higher performance. The gap between *SPARK* and the baselines increases in the zero-shot setting on the IBM-Rank-30K dataset. Here, augmentation with DPR, Flan-T5, and Llama-2 is consistently inferior, as is the *SPARK* augmentation of the single-encoder methods. In summary, *SPARK* effectively combines dual encoding and data augmentation for strong task accuracy and generalization.

Comparison of augmentation variants (Q2). On the in-domain task, the dual encoder performs

Model	Augmentation	GAQ Corpus (in-domain)						IBM-30K (ZS)	
		Cogency		Effectiveness		Reasonableness		WA	
		σ	ρ	σ	ρ	σ	ρ	σ	ρ
BERT	-	0.3480	0.3268	0.2804	0.2821	0.3285	0.3356	0.1989	0.1751
XLNet	-	0.1790	0.1673	0.2008	0.1930	0.1778	0.1847	0.1989	0.1816
Dual BERT	-	0.3685	0.3619	0.3082	0.3143	0.3694	0.3848	0.1766	0.1588
GPT-3.5	-	0.2879	0.3146	0.3585	0.3902	0.3561	0.4073	0.2698	0.2794
BERT	DPR	0.3215	0.3182	0.2728	0.2763	0.2821	0.3012	0.0706	0.0774
XLNet	DPR	0.2459	0.1778	0.2259	0.2271	0.2024	0.2160	0.1201	0.1254
Dual BERT	DPR	0.3536	0.3525	0.3227	0.3224	0.3396	0.3497	0.1571	0.1408
Dual BERT	Flan T5 XL (all)	0.3262	0.3268	0.2850	0.2920	0.3445	0.3564	0.1034	0.1241
Dual BERT	Llama-2 (7B) (all)	0.3468	0.3516	0.3296	0.3516	0.3269	0.3529	0.1673	0.1398
BERT	GPT-3.5 (all)	0.3418	0.3340	0.2863	0.3047	0.3459	0.3664	0.1510	0.1374
XLNet	GPT-3.5 (all)	0.2675	0.1930	0.2450	0.2060	0.2142	0.2046	0.1318	0.1352
Dual BERT	GPT-3.5 (feedback)	0.4184	0.4187	0.3925	0.4037	**0.4174**	0.4318	**0.2742**	**0.2854**
Dual BERT	GPT-3.5 (assumptions)	0.3699	0.3704	0.3881	0.3946	0.3681	0.3806	0.1728	0.1769
Dual BERT	GPT-3.5 (similar quality)	0.4059	0.4038	0.4150	0.4252	0.3545	0.3775	0.2629	0.2037
Dual BERT	GPT-3.5 (counter)	0.4030	0.4092	0.4048	0.4143	0.3989	0.4171	0.2086	0.2044
Dual BERT	**GPT-3.5 (all)**	**0.4242**	**0.4371**	**0.4513**	**0.4762**	0.4135	**0.4362**	0.2238	0.2293

Table 1: Performance of Dual-BERT model with augmentations applied compared to the baseline models. The performance of the model achieving the best scores per metric is **boldfaced** and the second best score is underlined.

Aug. / Metric	Validity	Inform.	Relevance
Feedback	4.87	3.80	4.96
Assumptions	4.95	4.91	4.97
Counter-arguments	4.93	4.94	4.99
Similar quality	4.40	4.40	4.62

Table 2: Validity, informativeness, and relevance scores of the augmentations for argument scoring.

the best when it has access to the information from the four *SPARK* augmentations simultaneously. Among them, feedback is most effective for predicting argument cogency and reasonableness as it exposes flaws that directly relate to these metrics. Meanwhile, contextualizing through similar-quality arguments is optimal for predicting effectiveness, which we attribute to illustrating the connection between quality scores and the argument structure, format, and wording. On out-of-domain data, the best performance is obtained by the feedback-augmented dual BERT, which even outperforms using all augmentations. This is illustrated in Table 3, where the first two arguments receive positive feedback, with space for improvement by further elaboration or addressing of alternatives, directing *SPARK* to increase its score. The third argument receives more critical feedback, causing *SPARK* to decrease its score. While GPT-3.5 is generally able to highlight the salient points of an argument and provide valid criticism, we also note an occasional bias of the model towards maintaining a neutral or positive argument stance (as in the case of the libertarianism argument).

Human judgment of augmentations (Q3). To validate the alignment of the augmentations with human utility, we performed a human study where we asked participants to score the validity, informativeness, and relevance of each augmentation strategy. The participants were asked to score 50 randomly sampled in-domain data points on these three metrics using a Likert scale of 1 (lowest) to 5 (highest). The results in Table 2 show that the augmentations are perceived by people as highly valid, informative, and relevant. Assumptions and counter-arguments were found to be consistently more valid, informative, and relevant than the other augmentations. Curiously, the participants judged the feedback informativeness to be lower, explaining that it often summarizes the argument instead of giving writing suggestions. This finding provides a cue that highly effective augmentations for models may not be perceived as informative by people. *SPARK* alleviates this issue by effectively combining the complementary augmentations and delegating the weighting of their utility to the model.

5 Conclusions

This paper enhances argument quality estimation models by providing contextualized feedback, inferred assumptions, and similar quality arguments or counter-arguments. We employ a dual-encoder Transformer to compare the argument and additional evidence effectively. Experimental results demonstrate that the best performance is achieved with the combination of all augmentations, indicating their complementary insights. *SPARK* outperforms single BERT, and XLNet in handling longer sequences, surpassing baselines and alternative augmentations in both in- and out-of-domain scenarios. Feedback augmentation is the single most effective augmentation strategy for models, despite being scored the least informative by human participants.

6 Limitations

The limitations in this paper primarily stem from the generative capabilities and hallucination tendencies of the LLM used for augmentation. For example, despite constraining the output format for the assumption augmentation task, the LLM still generates "No assumptions" as output after listing a set of valid assumptions. As future work, further augmentation studies must be performed to analyze and improve the prompts to minimize the misunderstanding and biases of the LLM.

Moreover, evaluating on a larger set of domains, languages besides English, and other argumentative tasks such as logical fallacy detection is an important next step in investigating the generalizability of our method. While in theory, *SPARK* can be directly applied to such tasks, it remains to be seen to which extent the current architecture and augmentation strategies will generalize to other tasks.

Finally, our analysis in this paper focuses on the comparison of methods and strategies, yet, we do not dive deep into the specific differences in performance across the three quality metrics which is an important future direction.

7 Ethics Statement

Using LLMs like GPT-3.5, with large and inaccessible pretraining corpora, can potentially lead to the amplification of biases in downstream argument quality ranking models. While we believe that these biases are sometimes necessary to successfully judge the quality of the argument, sampling assumptions from such models can lead to biased and unethical information being fed to (and amplified by) the model. A possible way to minimize this harmful knowledge for scenarios that involve sampling assumptions is by providing a fallback instruction in the prompt for the model to output "No assumptions"; in the present paper, we did not conduct specific studies to measure the impact of this fallback strategy because we did not see a significant impact on a small number of samples. Furthermore, to be able to compare with pre-existing baselines, we did not de-bias or anonymize the datasets provided but we strongly suggest that this should be considered wherever the *SPARK* method is deployed.

References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Alhindi, Tariq. 2023. Computational models of argument structure and argument quality for understanding misinformation.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530, Berlin, Germany. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Michael Fromm, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy

Faerman, and Thomas Seidl. 2022. Towards a holistic view on argument quality prediction.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *CoRR*, abs/1811.08008.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Gordon Harvey. 2009. A brief guide to the elements of the academic essay. *Harvard College Writing Program*.

Md Kamrul Hasan, James Spann, Masum Hasan, Md Saiful Islam, Kurtis Haut, Rada Mihalcea, and Ehsan Hoque. 2021. Hitting your MARQ: Multimodal ARgument quality assessment in long debate video. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6387–6397, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ioana Hulpus, Jonathan Kobbe, Christian Meilicke, Heiner Stuckenschmidt, Maria Becker, Juri Opitz, Vivi Nastase, and Anette Frank. 2019. Towards explaining natural language arguments with background knowledge. In *PROFILES/SEMEX@ ISWC*, pages 62–77.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021. Automated essay scoring using transformer models. *Psych*, 3(4):897–915.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yeti Mulyati, , and Daris Hadianto and. 2023. Enhancing argumentative writing via online peer feedback-based essay: A quasi-experiment study. *International Journal of Instruction*, 16(2):195–212.

Raymond S. Nickerson. 2020. *Biases, Misconceptions, and the Like*, page 208–262. Cambridge University Press.

OpenAI. 2022. Chatgpt. `https://openai.com/blog/chatgpt`. Accessed: April 30, 2023.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Dietrich Trautmann, Michael Fromm, Volker Tresp, Thomas Seidl, and Hinrich Schütze. 2020. Relational and fine-grained argument mining. *Datenbank-Spektrum*, 20(2):99–105.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. 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, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A Appendix

A.1 GPT 3.5 prompt templates

Feedback

The feedback on writing is sampled by considering both the topic and the argument related to the topic. To ensure brevity, we output the feedback in bullet point format. We follow the format below for sampling the feedback from the LLM:

Give concise writing feedback for the following argument in context with the topic, preferably in bullet points:
Topic: *topic*
Argument: *argument*.

Assumptions

Similar to feedback, assumptions are sampled in bullet point format to ensure brevity. Additionally, to constrain the hallucinations of the LLM, we restrict it to output "No assumptions" for the cases where it does not find assumptions or biases. We use the below prompt to sample this assumptions list:

Summarize the assumptions, if any, in the following argument in a bullet format otherwise return "No assumptions"
Topic: topic
Argument: argument.

Similar quality argument

To sample a similar quality argument, we use the following template:

Cogency Score: cogency score
Effectiveness Score: effectiveness score
Reasonableness Score: reasonableness score

Feedback	- The argument takes a clear position that cell phones cause distraction while driving. - The writer expresses strong emotions regarding the topic, which may influence the persuasiveness of the argument. - The argument does not provide any evidence or supporting details beyond the writer's personal experience and emotions. - The argument does not address whether cell phones cause distractions in other contexts beyond driving.
Assumptions	- Cell phones cause a distraction when driving. - Bluetooth or speaker use of cell phones can also be a distraction while driving. - People don't take driving as seriously as they should.
Similar quality argument	Yes, cell phones can be a major distraction to people. People are constantly checking their phones for notifications, messages, and other tasks that can take away from what they should be focusing on. This can lead to people not paying attention in class, not being productive at work, and not being present in conversations. Cell phones are an amazing invention, but they can also be a major distraction.
Counter-Argument	While it is true that distracted driving can be a major problem, it is unfair to blame cell phones as the sole cause of distraction. Many other factors can also contribute to distracted driving, such as eating, applying makeup, or even talking with passengers. Furthermore, cell phones can also be used in a responsible manner while driving, such as with the use of hands-free devices or by pulling over to take a call. Therefore, instead of demonizing cell phones, we should focus on educating drivers on safe driving practices and the dangers of distracted driving in general.

Table 3: Augmentation examples for the topic *Do cell phones cause a distraction to people?* with argument: *If you mean do they cause a distraction when driving the answer is a resounding YES!!! That includes bluetooth or speaker. Nothing makes me more angry when people don't take driving as seriously as they should. It is a major killer and if you don't follow the rules and focus you are putting other people in danger.*

Effectiveness Score	Generated similar quality argument to: Do you equate legality with morality?
1.0	Yes, I think that if the law says something is wrong, then it is wrong. People should follow the law and not do anything illegal.
2.0	No, I do not think that legality necessarily equates to morality. Laws can often be outdated, biased, or simply not reflective of what is right and wrong. For example, slavery used to be legal in many countries, yet it was obviously immoral. So legality and morality do not always align.
3.0	No, I do not equate legality with morality. Just because something is legal does not make it moral, and just because something is illegal does not make it immoral. Laws are made by people and can be changed, while morality is often determined by an individual's personal values and beliefs. Therefore, legality should not be used as a measure of morality.
4.0	No, I do not equate legality with morality. I believe that there are laws that should be followed in order to maintain the peace and order of a society, but there are also moral standards that must be adhered to for the same purpose. For example, it may be legal in some places to drive over the speed limit, but it is still morally wrong to do so.
5.0	I do not necessarily equate legality with morality. Legality is based on laws, which can be changed to reflect the morality of a society. Morality, in contrast, is based on principles and values that don't necessarily have to be enforced by law. For example, while it may be legal to drive over the speed limit, it is not necessarily moral to do so.

Table 4: Similar quality examples.

Topic: topic

We use ten samples in the few shot setting with two each from every integer ranking from 1-5 on the ranking scale for each metric. Finally, we prompt the LLM to generate the argument with respect to the cogency, effectiveness and robustness scores.

Counter-argument
The counter-argument is generated using the given argument and topic, and the following template:

Give a counter-argument for the following argument with respect to the Topic: topic

Argument: argument

A.2 Augmentation examples

Table 3 shows augmentation examples for one topic-argument pair. The topic questions if cell phones distract people and the argument agrees with it in context to distracted driving due to cell phones.

Feedback: The feedback discusses how the argument takes a clear position but appears overly emotional while answering, which may influence the persuasiveness of the argument. Additionally, the feedback regarding lack of evidence other than personal experience and the lack of discussion on

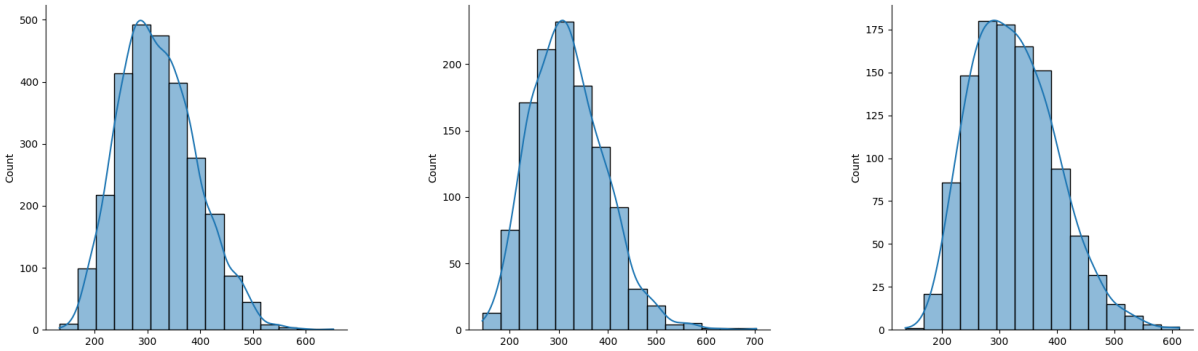

Figure 3: Distributions for augmentation lengths for the training, validation, and testing splits, respectively.

Dataset split	Dataset size	Minimum length	Maximum length	Mean length	Variance
Train	2746	133	652	320.27	5309.61
Validation	1177	144	703	318.55	5690.65
Test	1139	136	613	324.60	5467.65

Table 5: Distribution properties across splits for the concatenated augmentations.

phones causing a distraction beyond driving aim to help improve understanding while ensuring that the model pays careful attention to the topic in context.

Assumptions: The LLM lists several important assumptions in the proposed argument with respect to the topic. The first is the base assumption of the author's perspective, which is that cell phones cause distraction during driving. The LLM extracts the sentence *" people don't take driving as seriously as they should"* and labels it as an assumption because this is a faulty generalization to apply a general rule to all people, which is, in this case, people not taking driving seriously.

Similar quality argument: The generated similar quality argument tries to replicate the structural pattern of the given argument. The similar quality instance contains open-ended, long-winded sentences such as *"Cell phones are an amazing invention, but they can also be a major distraction"* which reduce the score of the argument, similar to the original argument. We also see that the LLM understands the ranking progression in a few shot setting and similar to the original argument, the similar quality argument also focuses solely on driving based cell phone distractions.

Counter argument: We notice that the LLM recognizes that the original argument only discusses distracted driving and so it only produces a counter argument of the stance that distracted driving is not the only cause for distraction. The response discusses the safe use of cell phones such as hands free, etc and advocates educating drivers on the effects of cell phone based distracted driving.

A.3 Impact of effectiveness score on GPT 3.5 outputs for similar quality arguments

Table 4 discusses the impact of effectiveness score on the generated similar quality argument. As can be seen in Table 4, the generated argument with an effectiveness score of 1.0 oversimplifies the relationship between legality and morality and treats related laws as fixed. Comparatively, the argument with an effectiveness score of 2.0 provides an example of slavery which enhances the effectiveness of the argument. Despite the addition of this example, the second argument lacks elaboration on why the example is immoral and fails to provide relevant evidence. The argument generated, given an effectiveness score of 3.0, recognizes that the law is not the sole arbiter of morality and that laws are subject to change. It does not only highlight the potential flaws in legal systems but also addresses the distinction between personal values and the law. However, this argument oversimplifies morality by implying that personal values and beliefs solely determine morality and lacks supporting evidence for the statement: *just because something is legal does not make it moral, and just because something is illegal does not make it immoral.*. The argument generated with an effectiveness score of 4.0 considers the coexistence of legal and moral standards. The addition of a specific example in this argu-

Table 6: Feedback examples for the out-of-domain dataset.

Topic	Argument	Score before augmentation	Score after augmentation	Ground truth score	Feedback augmentation
The use of public defenders should be mandatory	A centralized system of criminal defense would mean that all people would have access to the same standard of legal counsel, meaning that wealth and power can't be used to avoid justice.	0.628	0.675	0.828	- Clear and concise argument presented - Supports the idea of a centralized system of criminal defense - Highlights the importance of equal access to legal counsel - Addresses the issue of wealth and power influencing justice - Could benefit from further elaboration or evidence to strengthen the argument
We should ban algorithmic trading	algorithmic trading has been responsible for several mini market collapses, since computer systems lack the human sensitivity to look outside the stream of meaningless numbers to a wider context.	0.585	0.631	0.948	- Clear and concise argument presented - Provides specific examples to support argument - Could benefit from further elaboration on the potential consequences of mini market collapses - Could also benefit from addressing potential counterarguments or alternative solutions to the issue
We should adopt libertarianism	libertarianism is a justification for greed and exploitation	0.540	0.519	0.666	The argument is not specific enough about what adopting libertarianism entails - It assumes that libertarianism automatically leads to working together for a greater good and favoring the less well off, which is not necessarily true - The argument could benefit from providing concrete examples of how libertarianism would benefit society as a whole - It is unclear how giving freedom of choice would lead to greater societal benefits, and this point could be expanded upon.

ment adds concreteness and strengthens its persuasiveness. However, the argument can be further strengthened by acknowledging a broader range of situations where legality and morality may diverge. Finally, the argument ranked with the highest effectiveness score emphasizes the independence of morality from legal enforcement, which makes it even more persuasive. The contrasting comparison adds clarity to the flow of the argument and hence makes it better than all previously generated arguments.

A.4 Distribution analysis of augmentation lengths across splits

In this subsection, we conduct a distribution analysis on the augmentation input sizes to justify the use of the dual BERT architecture. Based on the findings presented in Table 5, the training split exhibits a minimum tokenized sequence length of 133 tokens, a maximum length of 652 tokens, and an average length of 320.27. The distribution of the training set, as presented by Figure 3, shows that only 0.5% (15 examples) of the training data exceeds BERT's token limit of 512 tokens.

The validation split has a minimum tokenized length of 144 tokens, a maximum length of 703 tokens, and an average length of 318.55 tokens. The testing split, on the other hand, has a minimum length of 136 tokens, a maximum length of 613 tokens, and an average length of 324.60 tokens. The percentage of data points in the validation and testing splits as seen in Figure 3 that exceed BERT's token limit are only 1.1% (13 examples) and 1.31% (15 examples), respectively.

Hence, we can conclude that the second BERT encoder tasked with embedding the augmentations is able to capture all the information in the augmentations without truncating the augmentations.

A.5 Impact of feedback

Table 6 shows the impact of providing feedback on the quality scores given by *SPARK* for three arguments.

A.6 Implementation and Human Evaluation Specifications

A.6.1 Training setup

We used the Hugging Face library (Wolf et al., 2020) for the training and inference of our models. The training of the dual BERT architecture was performed using 6xA5000 GPUs; each training run for five epochs with a batch size of 32 on each device and a learning rate of 5×10^{-3} takes approximately 30 minutes to converge. We found that the convergence was sensitive to the learning rate and smaller learning rate within the range of 1×10^{-4} to 5×10^{-3} was preferable for optimal results. The cosine scheduler provided the fastest convergence out of the possible linear, cosine, and polynomial scheduler options, and the best model was picked according to the highest F1 score on the validation split of the GAQCorpus. The inference for Llama-2 (7B) and Flan T5-XL (2.85B) was performed on 6xA5000 GPUs and took approximately 7 hours to complete. The GPT-3.5 experiments were repeated three times, and the average of the three runs is reported in Table 1. We maintained a temperature and top_p of 0.01 and 0.9, respectively, for all LLM experiments.

A.6.2 Human Evaluation

Our human study poses the following three targeted questions to the participants:

1. How **valid** is the information provided by the augmentation with respect to the background of the argument?

2. How **informative** is the augmentation for the task of argument quality analysis?

3. How **relevant** is the augmentation to help with the task of assessing the quality of the argument?

Our study was done with a sample of five computer science graduate students at the University of Southern California, and the results were averaged and reported in Table 2. To mitigate potential issues arising from sensitive arguments, we ensured that the questions were clearly understood and terms were explained, after which we obtained oral consent from each participant. The study was exempted from review by IRB (application UP-21-00443).