ArguMentor: Augmenting User Experiences with Counter-Perspectives

Priya Pitre[1], **Kurt Luther**[1],
[1]Virginia Tech
priyapitre@vt.edu, kluther@vt.edu

Abstract

Opinion pieces (or op-eds) can provide valuable perspectives, but they often represent only one side of a story, which can make readers susceptible to confirmation bias and echo chambers. Exposure to different perspectives can help readers overcome these obstacles and form more robust, nuanced views on important societal issues. We designed ArguMentor, a human-AI collaboration system that highlights claims in opinion pieces, identifies counter-arguments for them using a LLM, and generates a context-based summary of based on current events. It further enhances user understanding through additional features like a Q&A bot (that answers user questions pertaining to the text), DebateMe (an agent that users can argue any side of the piece with) and highlighting (where users can highlight a word or passage to get its definition or context). Our evaluation shows that participants can generate more arguments and counter-arguments and have, on average, have more moderate views after engaging with the system.

Introduction

Opinion pieces form a significant part of our lives. An opinion piece is an article where the writer expresses their opinion on political, social, and societal issues. These often have one characteristic in common: they only represent one side of any story. Coppock, Ekins, and Kirby (2018) find that op-eds are persuasive to both the general public and elites, influencing opinions across society on major issues. Druckman (2005) shows that op-eds can inform major election decisions amongst voters as well. Considering the widespread popularity and significant impact of opinion pieces in our lives, it becomes crucial to address the potential challenges that may emerge while engaging with such content.

First, humans are susceptible to confirmation bias (Kaanders et al. 2022; Suzuki and Yamamoto 2021). Most people struggle with confirmation bias, whether they consciously realize it or not (Nickerson 1998). When reading opinion pieces, people often only read content that aligns with their preconceived notions. People naturally gravitate towards content that aligns with their interests and beliefs. However, the concern arises when such one-sided content is all that people consume, they may lack more complete information that would lead them to reach a different conclusion.

Second, online information, such as news articles and posts are often targeted to exploit biases and create divides, driving more clicks and furthering specific agendas (**?**). As a result, it becomes nearly impossible to consume unbiased content, as most information is polarized. These factors can lead to the formation of echo chambers within society, driving conversations away from consensus-building discussions.

To combat both idle reading and media bias, researchers have proposed several paper support tools. Idle reading support tools (Fok et al. 2022; August et al. 2023; Kang et al. 2023) support highlighting, critical thinking, question answering, etc for increasing user attention span and understanding of the text. Several papers and websites have tackled media bias by providing multiple perspectives (Park et al. 2009; Perez et al. 2020; Hamborg et al. 2020), guiding users to the same article written from the other side (AllSides Technologies 2012; Ground News Technologies 2024), etc. However, these approaches require the user to take the additional effort to read another article, which users don't always have the time or motivation for. Moreover, there is a delayed effect, where the user could already be influenced by one article before reading the next one. We argue that users benefit from seeing counter-arguments next to the arguments in real time as they are reading an article.

To address these issues, we present ArguMentor, a human-AI collaboration system that enriches the reading experience of opinion pieces by highlighting claims and generating counter-arguments to help users formulate more robust, nuanced views on important societal issues. ArguMentor provides two types of support: passive and active. The system provides passive support by automatically highlighting main claims in the original text, generating counter-arguments for them, and providing a context based summary for the overall text. ArguMentor provides active support by allowing users to interact with the system by accessing the Q&A bot (that answers user questions pertaining to the text), a DebateMe feature (an agent that users can argue any side of the piece with), and a highlighting trigger window (where users can highlight a word or passage for its definition or context). We derived these features from a preliminary study of news readers and debaters, which illuminated key challenges and potential features and informed our design goals. We evaluated ArguMentor via a within-subjects experiment (N=24). We found that, compared to the baseline, ArguMentor was able to help participants generate more claims, boost their memory for those claims, and even changed participants' opinions. In summary, the contributions of this paper are as follows:

- We propose an opinion article reading tool ArguMentor that augments user experience with AI-generated counter-arguments, a debate bot, and a question answering bot.
- We demonstrate the effectiveness of ArguMentor via a within-subjects experiment with 24 participants.
- We provide insights and design considerations for future similar tools.

Related Work

Paper Reading Support Tools

HCI researchers have proposed a variety of software tools and technologies to support reading comprehension and critique of academic papers. To support reading comprehension, Kim et al. (2018) creates an interactive document reader that links the text with its corresponding table cells automatically, which can reduce split attention and facilitate reading. August et al. (2023) created Paper Plain, a system that utilizes NLP techniques to enhance understanding of medical papers. Fok et al. (2022) created a skimming tool that highlights specific parts of the text in different colors to guide the reader's attention, and enable them to read more efficiently. (Kang et al. 2023) makes the process of finding other papers related to the current paper the user is reading easy by creating a citation graph and threading and summarizing their content using GPT-4. Other recent work (Rachatasumrit et al. 2022; Kang et al. 2022) adds on to this process by automatically highlights important citations, and provides commentary for them based on these papers, and creates a social network for various papers.

To support close reading and critique of written work, Tan et al. (2016) propose a web-based collaborative platform that supports peer interactions and provides feedback for both students and teachers to engage in critical paper reading together. Crebot (Peng et al. 2022) asks users critical questions as they are reading the passage to further their understanding of it. Yuan et al. (2023) uses text summarization techniques and template based questions to help users raise critical thoughts.

Our work builds upon these advancements by incorporating key features such as highlighting important points, summarizing content for better user understanding, and developing a question-answering framework. By adapting these established techniques to the domain of reading opinion articles, we aim to address the unique challenges associated with engaging critically with one-sided narratives, promoting a more balanced and comprehensive understanding among readers.

News Reading Support Tools

HCI and News: Extensive research has been conducted to enhance news consumption through innovative technologies and interfaces.

Laban et al. (2023) design and evaluate news reading interfaces that incorporate discord questions to reveal coverage diversity. Chen et al. (2023) developed Marvista that employs Natural Language Processing (NLP) techniques like abstractive summarization to provide text-specific assistance when users are reading online articles. Their main evaluation study showed that Marvista helps users better comprehend the article. Nguyen et al. (2018) blends information retrieval and human knowledge to create a fact checking portal that aids human fact checkers.

Combating Media Bias using HCI tools: NewsCube (Park et al. 2009) provides readers with multiple perspectives on the same news, by showing them the article written by multiple sides. AllSides Technologies (2012) also does a similar thing by ranking media outlets, and showing a right, left and centrist perspective on the news. Perez et al. (2020) propose a browser extension that presents different perspectives by recommending articles relevant to the current topic. Munson and Resnick (2010) evaluate the extent to which highlighting user agreeable terms within text or showing them first has an impact on their opinion. Hamborg et al. (2020) evaluates word choice bias in media by highlighting trigger words like "terrorists" etc and indicating their positive or negative sentiment.

Our work differs from these efforts in two key ways. First, ArguMentor allows the user to simultaneously read the article and its counter-argument instead of referring them to another source. We expect that this quicker presentation has a beneficial effect on the user in that they are more likely to read the counters, and less likely to be influenced by the article in the first place. Second, ArguMentor refutes the argument from the article using an LLM-generated counter, instead of another (potentially biased) human-written article from an opposing viewpoint. This allows us to provide a succinct, relevant, contextualized counter-argument rather than directing the reader away from the current opinion piece in order to read additional opinion pieces.

Integration of LLMs to persuade users

The use of Large Language Models (LLMs) for tasks related to persuasiveness has been explored in several other contexts. Hyben et al. (2023) tests LLMs and fine-tuned models for claim detection to tackle things like misinformation and spread of bias. Khan et al. (2024) shows that debating with persuasive LLMs leads to truthful answers and shows that debate with an LLM is a good way to resolve conflict in cases where ground truth is unavailable. Breum et al. (2023) shows that chatbots and LLM agents can generate powerful and persuasive arguments, and how they can play an important role in online discussions. Argyle et al. (2023) uses chatbots to show that online political conversations can be improved with an AI assistant's suggestions. Karinshak et al. (2023) demonstrate that AI can be used to create effective public health messages, and that people are often persuaded by messages created by LLMs.

These prior works lay the groundwork showing that LLMs have the potential to be persuasive and change users' opinions. We use LLMs in a similar way (by leveraging them as chatbots and prompting agents), but adapt and extend these ideas to the novel context of an interactive system supporting engagement with opinion pieces.

Preliminary Study

To support the creation of this system, we conduct an initial survey to gauge the challenges people face when reading opinion pieces and to better understand user attitudes and preferences towards potential design features.

Potential List of Features

To brainstorm the potential features a reading experience aiding system could have, we reviewed the aforementioned recent literature in paper reading tools (e.g., (Fok et al. 2022; Kang et al. 2022)), along with asking users for their feedback on a potential list of features. The final list of potential features that is presented to participants of the survey is shown in Table 1.

Survey Method

We recruited 21 participants (primarily young, educated, English-speaking news readers) to complete our survey. The questions focused on their demographics and news reading habits, information about their interaction with opinion pieces, and design ideas for potential technological support. Details of the survey protocol and participant backgrounds can be found in the Supplementary Materials.

Findings

Challenges of Reading Opinion Pieces

- **C1: Difficulty Focusing and Completing the Articles:** A majority of the participants indicated that they struggle with long opinion pieces because it is difficult to stay engaged in a long piece. P3 writes, "Opinion pieces are often really long, and I find myself reading a passage, dozing off, then having to re-read it."
- **C2: Confirmation Bias:** Consistent with prior work, a few participants note that they struggle with confirmation bias, where they much prefer reading articles that support their point of view, and avoid reading articles from the other side altogether. P1 notes, "Sometimes its difficult to avoid the tunnel vision that you experience when arguing for or against a topic in a specific way."
- **C3: Difficulty Imagining Counter-arguments:** A few of our participants also noted that they have a hard time thinking about what the other side might say, when they are invested in reading an opinion piece. They indicate that they often have strong opinions about the current piece, one way or another, and can't easily dismiss those thoughts. When surveyed on whether they go back to researching the other side once they're done reading, a majority of these participants indicated that they don't, and often the opinion from the article sticks with them. P1 responded, "It becomes difficult to think of creative ways to defend/oppose the argument from the templated way in which you are used to doing so."

Usefulness of potential features Participants were asked to rate potential features on a scale of 1–5 (5 = most useful). Table 1 shows the results for this part of the survey.

Open-ended questions Our participants also responded to open-ended questions about other potential features they would like to see. P18 suggested that they would like to see the main claims from the passage highlighted right next to the counterclaim, which is also highlighted in the same color. A few participants suggested that the "DebateMe" feature should take feedback about persuasion from the user, and modify its arguments accordingly. Similarly, some participants suggested that the counterarguments should also take user feedback for whether they are persuasive. The final system was designed based on feedback from this initial survey, and will be described in the following section.

Potential Features	Mean	StdDev
Highlight main arguments	**4.33**	**0.79**
Counter-arguments	**4.24**	**0.88**
Current events context	**3.95**	**0.86**
Summary	3.89	1.08
Left vs Right comparison	3.71	1.10
DebateMe	**4.19**	**0.98**
User style adoption	**3.82**	**0.97**
Highlight for context	**3.80**	**0.93**
Q&A	**3.71**	**0.85**

Table 1: Feature Survey Results

System Description

Based on initial feedback and related research, we developed the final architecture for the system, as shown in Figure 1 The system is divided into 2 stages: passive interaction and active interaction. Informed by the preliminary study and our literature review, we created ArguMentor, a human-AI collaboration system that enriches the reading experience of opinion pieces. In this section, we present a user scenario, followed by a description of the back-end architecture.

User Scenario

In this user scenario, we describe how John, a frequent news reader, would use this system. John begins by providing an opinion piece about monetary policy to ArguMentor that he wants to read. Next, he will see the op-ed text with claims highlighted on the left, along with counter-arguments for those claims on the right. He has limited time to read right now, so he clicks on the "get extra context" button, which generates a summary and provides broader context from online information. Then he skims through the counter-argument summaries.

Later, John uploads a second opinion piece on the Trump indictment when he has more time to read. He's less familiar with this topic and isn't sure where he stands on the issue. While reading, John encounters a legal term, "statute of limitations," so he uses the Highlight feature to select that text and request a contextualized explanation. Then he uses the Q&A feature to type in a question about the demographics

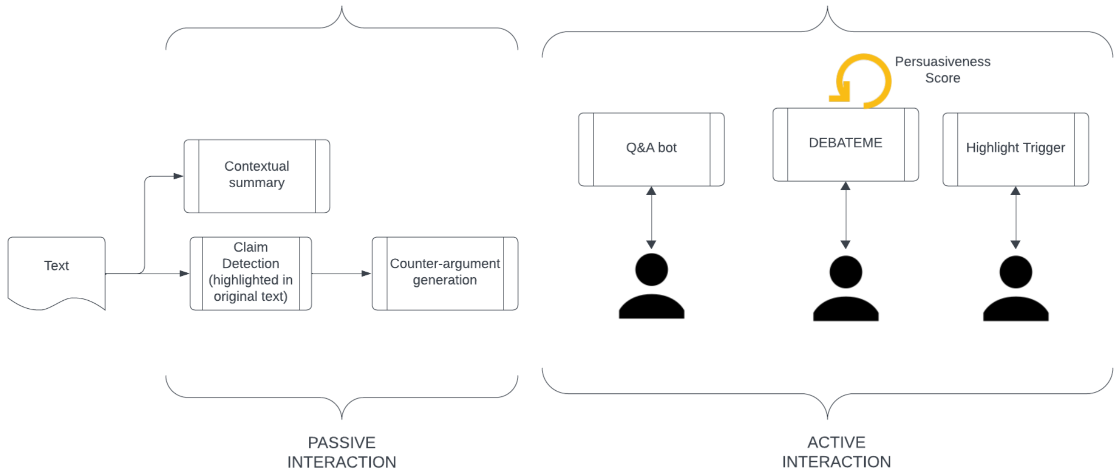

Figure 1: Architecture Diagram

of the cities mentioned in the article. Finally, he feels comfortable he understands the author's key points. He activates the DebateMe feature to express a key belief he shares with the author; the bot plays Devil's Advocate and provides an opposing viewpoint in real time.

Having completed the interaction, John now has a better understanding of not only the opinion piece, but also of where he stands on this topic. He can also argue for it better in the future. In summary, John can passively or actively interact with the system to ensure that he is not influenced by just this article, and knows the full context of the piece.

System Architecture

ArguMentor is deployed on Vercel (Inc. 2015), and Mouseflow (Inc. 2011) is used on the website to track user activity and record sessions.

Passive Support: There are three forms of passive reading support: context of the article, highlighted main claims in the article, and counter arguments for those claims.

1. **Summary/Context**: Upon clicking a button that says "Get Additional Context", users can get a neutral context about the article from the internet. This is done using SerpAPI (Inc. 2017). GPT-3's knowledge is limited to 2021, as of this paper, and hence getting context about recent articles is not possible just using GPT-3 API. Using SerpAPI and its "zero-shot-react-description" agent can be used to get direct google search results. The prompt passes the title of the article and asks to summarize the context of this issue. This achieves two purposes: gives the readers a quick summary, and ensures that the summary is not biased, but rather based on the context of the passage. Users in our preliminary survey indicated that they preferred this summary over a summary of the article, hence it has been implemented.
2. **Claim detection**: Claims are automatically detected within the text and highlighted. A fine-tuned model is used for claim detection. The GPT-3.5-turbo model is fine-tuned using the IBM-30K claims dataset (Aharoni et al. 2014). First, we structured the dataset according to the fine-tuning guidelines of OpenAI, where the system instruction was to return claims as it is, and user and assistant instructions were examples from the dataset. Only 11 instances are able to produce a good result for fine-tuning a large model like GPT3.5 (OpenAI 2022). Once the model is able to return claims from the passage, the claims are matched to the original text using REGEX. An instance of GPT-3.5 cannot be used by itself because we need to return an exact match so it can be highlighted- this is something a fine-tuned model achieves significantly better. The match is then highlighted in yellow. The main challenge in this part was to create the fine-tuning dataset such that it would account for a diverse range of news articles of various lengths.
3. **Counter-argument generation**: This is a basic instance of in-context few-shot learning (Brown et al. 2020). A few examples are provided in the prompt to show an instance of GPT 3.5 what a potential counter-argument could look like, and it is asked to generate counter-arguments for the claims that are passed to it in a list. This is then displayed on the right hand side of the page. The users can click on "Expand" to see the whole counter-argument if they are interested in it. Although most times the contexts are efficient and refute the claim, sometimes they can be vague, especially if the claim is about a recent event that the LLM is not trained on.

Active Support: Active support (interaction) comes primarily in three forms: Q&A, DebateMe, and highlighting.

1. **Q&A**: There were two options for creating a Q&A agent: a prompting method like counter-argument generation, or a RAG architecture. A RAG architecture is less likely

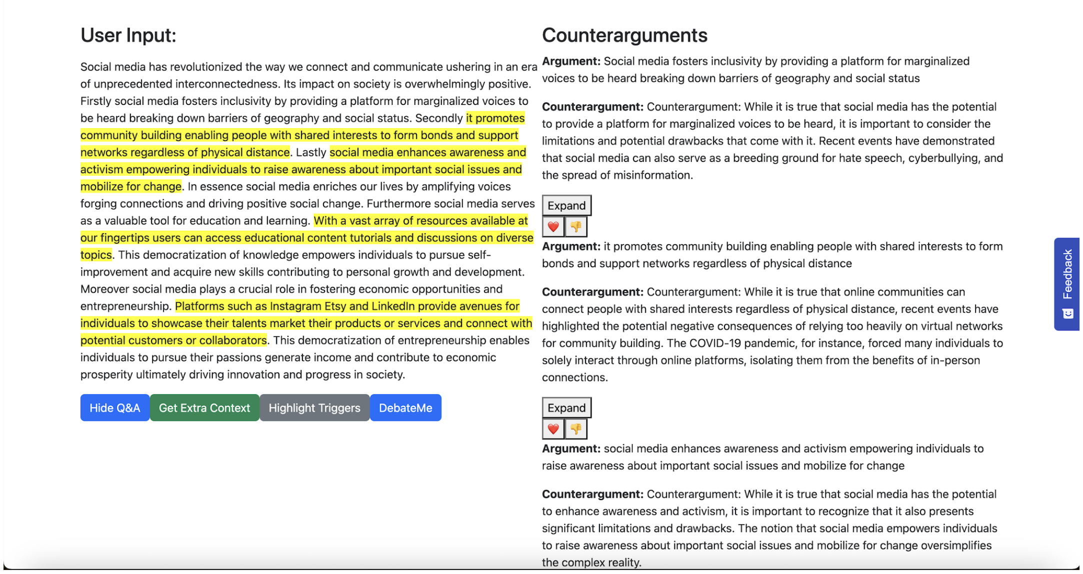

Figure 2: ArguMentor System Screenshot

to hallucinate, provides more relevant responses, and reduces biased responses (Gao et al. 2024). Hence, the latter was chosen for this architecture. OpenAI embeddings were used to convert the user query and the document that the user has uploaded into a vector space, and store it in the Chroma vector store. A button was created that initiated a chatbot where users could type their query and get efficient responses from the bot. History of the conversation is also shown on screen.

2. **DebateMe**: This is implemented using LangChain's conversation chain which enables chains of conversation. Similar to **?**, the user can enter their argument, and the bot will debate the other side. This is implemented using prompting — the LLM is asked to debate the other side to the user's argument, and be as brief and persuasive as possible. The user is then given an option to click on a "thumbs up" or "thumbs down" button. If the user clicks on "thumbs down", the LLM will find another way to persuade the user. This is done to ensure that the user is satisfied with the response, and can report any problematic content to the developers. This is also done using in-context prompting.
3. **Highlighting window**: A window is enabled where the user can highlight specific parts of the text to get more information about it, including definitions, counter-arguments, additional information, etc. The text selected by the user is sent along with the following prompt: "If it's one word, provide its definition. If it's more than that, use your knowledge to give additional context on the text".

Evaluation

To investigate the benefits of ArguMentor for users reading opinion pieces, we conducted a within-subjects experiment with 24 participants. In this experimental design, each participant read some opinion articles using ArguMentor (experimental condition) and others without the system (control condition), i.e., typical reading without any intervention.

Task and Procedure:

The experiment is conducted in several parts. The entire procedure is shown in Figure 3.

First, participants give their stance on the main topic of six articles on a scale of 1–5 (1=strongly disagree, 5=strongly agree). Then, participants are given a polarizing article from the right in the baseline (no ArguMentor) condition. A polarizing article is defined as one that is based on one of the major talking points of the political right or left (e.g., abortion, immigration), where most people would have a strong reaction to the article one way or another. Although the terms "left", "right" and "neutral" as used in this paper refer to American politics, the articles don't require the users to know about American politics to be able to participate. They are then asked to fill a survey which asks for how many claims they can spot, and counter-arguments they can think of. They are also asked to answer some basic attention checking questions. This is repeated for left-leaning articles, and finally a neutral article. A break is given after this procedure.

Then, participants repeat this process for three additional articles (representing left, right, and neutral perspectives) in the experimental (ArguMentor) condition, including the post-surveys. TA final, exit survey consisting of subjective experience and system feedback questions is given in the end. Participants can have as much time with each article and the system as they wish. They are instructed to not return to the article or the system while filling the survey, and answer subjective questions as honestly as possible.

The document selection is counterbalanced to reduce order effects; i.e., Participant 1 gets Documents A–C in the

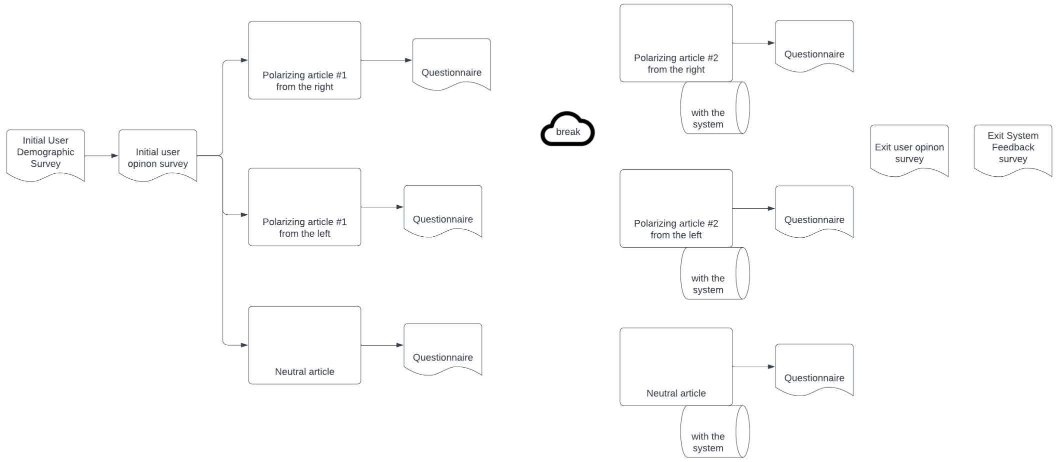

Figure 3: Experiment Design and Procedure

control condition and D–F in the experimental condition, Participant 2 gets Documents D–F in the control condition and A–C in the experimental condition, and so on. However, all participants experienced the experimental condition after the control condition.

The articles are chosen from AllSides Technologies (2012), which collates daily top news stories from the political left, right, and center. To create a spectrum of reading comprehension issues, the article content is tweaked using ChatGPT to make one article more challenging to read (adding difficult language, jargon, etc), one medium difficulty, and one easy to read. The six articles are included in the Supplementary Materials.

Evaluation Metrics:

We measured participants' outcomes, subjective experience, and processes to see how interaction with ArguMentor compares to the baseline.

RQ1: Outcomes We measured participants' outcomes in two areas: finding claims and thinking of counter-arguments for those claims. As our preliminary study indicated, users have a hard time doing both of these tasks, so we want to see if the system improves these skills. We count the number of claims/counterarguments participants submit in each post survey. Our hypothesis is that use of ArguMentor will increase the number of both.

RQ2: Process We observe the process by seeing how long the participants spends on which feature, and how that contributes to the overall number of arguments. We also measure the number of arguments the participants come up with and divide that by the time they spend on the baseline or system to normalize it, and see the effect the system has on creating any additional burden for the user.

If the participants can come up with an equal number of arguments for both, but the system takes them twice as long to use, it might suggest that the system is burdensome to use; whereas the benefits are clearer if the participant can come up with more counter arguments after using the system for more or less an equal time than just reading the article. However, it is important to consider that in contexts where efficiency is not a priority, such as leisure reading, taking more time might be acceptable if the system offers a better subjective experience. We also want to measure how the user spends time on the system– which features they use, for how long, etc.

RQ3: Subjective Experience We measure the participant's subjective experience of the system on two metrics: how they feel about the system, and how they feel about the article after using the system. Specifically, the survey captures how strongly the participant felt about a topic without the system (on a scale of 1–5), and then how they feel about it after interacting with the system. This way we can test if the counter-arguments and the system has an impact on participants' perception of the topic. After that, we can adopt the technology acceptance model as well as a survey to take feedback on the system.

Participants:

We recruited 24 participants from our university and an online quizzing group. A majority speak English as their first language, read the news regularly, and come from a STEM background. Half identified as men; the others as women. All have at least a master's degree in formal education.

Results

We perform the Mann–Whitney U test (Mann and Whitney 1947) to compare the difference with and without the system for all research questions. The Mann–Whitney U is a non-parametric test commonly used to compare differences between independent conditions especially when the data normality is violated, as confirmed in our cases. In all U tests, we set the significance level at the standard threshold of a=0.05.

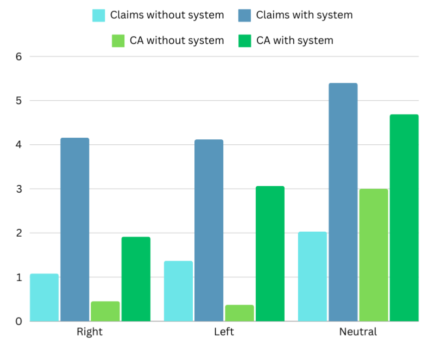

Figure 4: Performance: Number of Claims and Counter-Arguments with and without ArguMentor.

All 24 participants scored 100% on all the attention-checking questions.

RQ1: Outcome Results

Number of claims: Figure 4 shows the number of claims increase by 3, 4, and 5 times when the users use the system for neutral, right-leaning and left-leaning articles, respectively. These differences were significant (p=0.001, 0.0003, 0.0001). The overall word count also increased by over 200% for the claims.

Number of counter-arguments: Similar to claims, the number of counter-arguments users were able to generate increased significantly, ranging from a 0.6x increase (neutral) to a 6x increase (left-leaning) (p=0.03, 0.002, 0.0001 for neutral, right- and left-leaning, respectively). Figure 4 shows the number of counter-arguments with and without the system. The overall word count for the counter-arguments again increased by over 200% over the baseline. Detailed diagrams are in the Supplemental Materials.

RQ2: Process Results

Claims and Counter-arguments per minute: Figure 5 shows how claims per minute increase for all types of articles when participants use ArguMentor. On the other hand, counter-arguments per minute goes down for neutral articles, and goes up for left- and right-leaning articles. Participants spend more time and generate fewer counter-arguments when reading neutral articles with ArguMentor compared to reading without it.

Time spent on system by activity: Figure 6 shows how long participants spend on which feature in the system. This data is acquired using Mouseflow logs. Over half the time was spent on claims and counter-arguments, while the Q&A and DebateMe features were least used with less than 10% of overall time spent. Interestingly, these results align with the preliminary study's survey results in Table 1, where participants used main claims and counter-arguments the most.

RQ3: Subjective Experience Results

Change in views: Table 2 shows how user views changed for right-leaning, left-leaning and neutral articles before and after the system. As other studies have previously shown, political views are more entrenched, and harder to change, which is reflected in this study as well. The change in views for neutral articles was significant ($p<0.05$), whereas left and right articles were not.

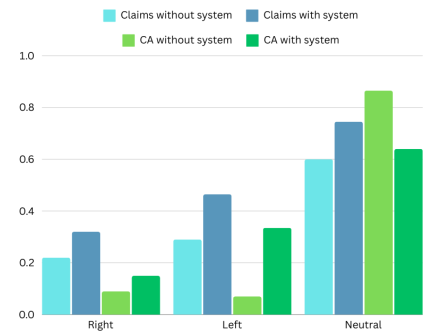

Figure 5: Process: Claims and Counter-arguments per minute with and without Argumentor

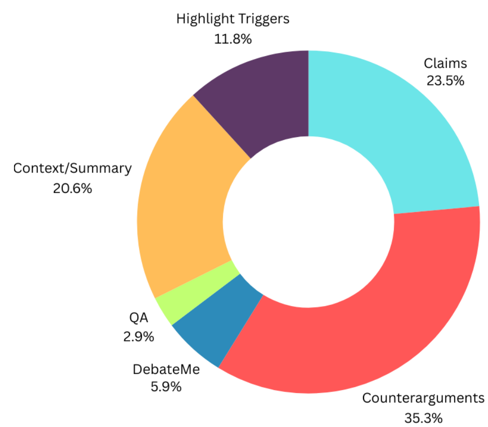

Figure 6: Time Spent using ArguMentor features.

System experience: Table 3 shows how the users perceived the system and its various features. The system scored high for helpfulness and ease of use. When asked in detail, users consistently reported DebateMe or Q&A features being the least helpful, consistent with the time spent analysis. Participants cited lack of interest in the articles, and the process being too time consuming as the primary reasons for not using these features. However, participants who found an article particularly interesting enjoyed the DebateMe Feature.

Discussion

In this paper, we propose ArguMentor, a system designed to help users gain a more nuanced, balanced understanding of news opinion pieces. We note that users often encounter one perspective when reading news articles, and could benefit from having access to counter-arguments. The larger goal of our system is not necessarily to change user opinions, but to give users quick, easy access to multiple viewpoints while reading an opinion piece, helping them form a more

Political Side	Article Topic	p-value
R	Large corporations should pay a much higher tax than the rest of us.	0.58
R	Minimum wage is a good idea	0.75
L	Sometimes, violence is the only way to protest.	0.21
L	Identity politics is an important idea for social movements to propagate their views	0.12
N	Social media does more good than harm	0.03*
N	Remote work is a net positive.	0.02*

Table 2: Subjective Opinions Results: Change in Views. * = difference is significant.

Factor	Mean	SD
Ease of Use	4.71	0.46
Motivation to Use	4.50	0.51
Helpfulness	5.00	0.00
Frustration (inv.)	4.13	0.68
Mental Demand (inv.)	3.3	1.07
Recommend to friend?	3.58	0.97
Daily Use	2.88	0.90
Persuasiveness	4.13	0.90

Table 3: Subjective Results: System Feedback

informed, robust, and nuanced perspective.

Addressing RQ1, we can see from the outcomes results that participants identified a significantly higher number of claims and counter-arguments with the help of the system. Without the system, most participants could only identify one main claim within the article. However, the system systematically highlights all claims throughout the article and provides them with specific counters, which can help users holistically understand the article. Participants were instructed not to return to the system or article when filling the survey, but were still able to write 4x as many points after using ArguMentor, showing that the system of highlighting also helps with memory. This is consistent with other HCI systems which use highlighting to boost memory (e.g., (Fok et al. 2022)).

In case of RQ2, we can see that the additional time participants spent on the system often led to a higher number of claims, and counter-arguments. The only exception was common knowledge topics like remote work and social media, where people already were able to come up counter-arguments without the system. However, users are able to generate longer counters with the help of the system (as indicated by a higher average word count), indicating higher confidence and potentially higher quality of counters. Beyond opinion pieces, news articles, especially about economics, science, politics, etc., can be complex, jargon-heavy, and unintentionally biased, indicating places where our system can provide additional value.

For RQ3, we found that ArguMentor led to significant changes in participants' views on topics, but only for neutral articles. It may be that neutral articles provided more viewpoints and opportunities for exploring counter-arguments and alternative perspectives, or that the topics of these articles (e.g., remote work) were less polarizing. Similar to other works in the field, we speculate that participants didn't change views for other articles because political views are more ingrained, and harder to change.

When the topic/article was particularly complicated, users used the highlight feature (and benefited from it) a lot more. The right-aligned economics article on taxation, which was naturally harder and jargon-heavier than others, showed an increased usage of highlight features, etc- and users later reported finding those features useful as well. Similarly, even though DebateMe was used less overall, users who were particularly passionate about a news topic used it (and found it useful) a lot more for that article. Users also reported finding the thumbs up and down features useful, as they were able to get several different arguments quite easily.

The results from this system can be used to present news in a different way to users. Journalists working for news companies could take inspiration from ArguMentor to make their straight news as unbiased as possible, whereas op-ed writers could use it to identify and address likely opposing viewpoitns. On social media platforms, AI and LLMs can also be used to automatically generate counter-arguments (like warnings) under some posts to educate and engage users. Humans benefit from exposure to multiple perspectives to be able to make better independent decisions, and ArguMentor's passive and active support features offer an important step towards that goal.

Limitations and Future Work

The goal of the system is to reduce the bias that articles bring in. However, since we are using LLMs to generate counter-arguments and for chatbots, it is possible that they are either not as persuasive, not as relevant, or can carry more bias or misinformation compared to those written by expert humans. We have mitigated this by adding certain constraints to the prompt, and using models that have guardrails (like GPT-3), and using a RAG system rather than simply prompting wherever applicable. However, there is still a possibility that the model response has bias, and it should be monitored continuously.

Additionally, all participants in the study were homogeneous in terms of high formal education and enthusiasm for news. Consequently, their willingness to read opinion pieces, consider multiple perspectives, and engage with system features may differ from a broader cross section of the news-reading public. However, our preliminary study indicated that even this participant pool could find it difficult to carefully read and focus on op-eds. Future work should extend this evaluation to a broader demographic pool.

References

Aharoni, E.; Polnarov, A.; Lavee, T.; Hershcovich, D.; Levy, R.; Rinott, R.; Gutfreund, D.; and Slonim, N. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In Green, N.; Ashley, K.; Litman, D.; Reed, C.; and Walker, V., eds., *Proceedings of the First Workshop on Argumentation Mining*, 64–68. Baltimore, Maryland: Association for Computational Linguistics.

AllSides Technologies, I. 2012. AllSides.

Argyle, L. P.; Bail, C. A.; Busby, E. C.; Gubler, J. R.; Howe, T.; Rytting, C.; Sorensen, T.; and Wingate, D. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41): e2311627120.

August, T.; Wang, L. L.; Bragg, J.; Hearst, M. A.; Head, A.; and Lo, K. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.*, 30(5).

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2023. The Persuasive Power of Large Language Models. arXiv:2312.15523.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, X. A.; Wu, C.-S.; Murakhovs'ka, L.; Laban, P.; Niu, T.; Liu, W.; and Xiong, C. 2023. Marvista: Exploring the Design of a Human-AI Collaborative News Reading Tool. arXiv:2207.08401.

Coppock, A.; Ekins, E.; and Kirby, D. 2018. The Long-lasting Effects of Newspaper Op-Eds on Public Opinion. *Quarterly Journal of Political Science*, 13(1): 59–87.

Druckman, J. N. 2005. Media Matter: How Newspapers and Television News Cover Campaigns and Influence Voters. *Political Communication*, 22(4): 463–481.

Fok, R.; Kambhamettu, H.; Soldaini, L.; Bragg, J.; Lo, K.; Head, A.; Hearst, M. A.; and Weld, D. S. 2022. Scim: Intelligent Skimming Support for Scientific Papers. *Proceedings of the 28th International Conference on Intelligent User Interfaces*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Ground News Technologies, I. 2024. GroundNews.

Hamborg, F.; Zhukova, A.; Donnay, K.; and Gipp, B. 2020. Newsalyze: Enabling News Consumers to Understand Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, 455–456. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375856.

Hyben, M.; Kula, S.; Srba, I.; Moro, R.; and Simko, J. 2023. Is it indeed bigger better? The comprehensive study of claim detection LMs applied for disinformation tackling. arXiv:2311.06121.

Inc., M. 2011. Mouseflow.

Inc., S. 2017. SerpAPI.

Inc., V. 2015. Vercel.

Kaanders, P.; Sepulveda, P.; Folke, T.; Ortoleva, P.; and De Martino, B. 2022. Humans actively sample evidence to support prior beliefs. *eLife*, 11: e71768.

Kang, H. B.; Kocielnik, R.; Head, A.; Yang, J.; Latzke, M.; Kittur, A.; Weld, D. S.; Downey, D.; and Bragg, J. 2022. From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.

Kang, H. B.; Wu, T.; Chang, J. C.; and Kittur, A. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. ACM.

Karinshak, E.; Liu, S. X.; Park, J. S.; and Hancock, J. T. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).

Khan, A.; Hughes, J.; Valentine, D.; Ruis, L.; Sachan, K.; Radhakrishnan, A.; Grefenstette, E.; Bowman, S. R.; Rocktäschel, T.; and Perez, E. 2024. Debating with More Persuasive LLMs Leads to More Truthful Answers. arXiv:2402.06782.

Kim, D. H.; Hoque, E.; Kim, J.; and Agrawala, M. 2018. Facilitating Document Reading by Linking Text and Tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, 423–434. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359481.

Laban, P.; Wu, C.-S.; Murakhovs'ka, L.; Chen, X. A.; and Xiong, C. 2023. Designing and Evaluating Interfaces that Highlight News Coverage Diversity Using Discord Questions. arXiv:2302.08997.

Mann, H. B.; and Whitney, D. R. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1): 50 – 60.

Munson, S. A.; and Resnick, P. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, 1457–1466. New York, NY, USA: Association for Computing Machinery. ISBN 9781605589299.

Nguyen, A. T.; Kharosekar, A.; Krishnan, S.; Krishnan, S.; Tate, E.; Wallace, B. C.; and Lease, M. 2018. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 31st Annual*

ACM Symposium on User Interface Software and Technology, UIST '18, 189–199. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359481.

Nickerson, R. S. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2): 175–220.

OpenAI. 2022. OpenAI.

Park, S.; Kang, S.; Chung, S.; and Song, J. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, 443–452. New York, NY, USA: Association for Computing Machinery. ISBN 9781605582467.

Peng, Z.; Liu, Y.; Zhou, H.; Xu, Z.; and Ma, X. 2022. CReBot: Exploring interactive question prompts for critical paper reading. *Int. J. Hum.-Comput. Stud.*, 167(C).

Perez, E. B.; King, J.; Watanabe, Y. H.; and Chen, X. A. 2020. Counterweight: Diversifying News Consumption. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, 132–134. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375153.

Rachatasumrit, N.; Bragg, J.; Zhang, A. X.; and Weld, D. S. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *27th International Conference on Intelligent User Interfaces*, IUI '22, 707–719. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391443.

Suzuki, M.; and Yamamoto, Y. 2021. Characterizing the Influence of Confirmation Bias on Web Search Behavior. *Frontiers in Psychology*, 12.

Tan, J. P.-L.; Yang, S.; Koh, E.; and Jonathan, C. 2016. Fostering 21st century literacies through a collaborative critical reading and learning analytics environment: user-perceived benefits and problematics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, 430–434. New York, NY, USA: Association for Computing Machinery. ISBN 9781450341905.

Yuan, K.; Lin, H.; Cao, S.; Peng, Z.; Guo, Q.; and Ma, X. 2023. CriTrainer: An Adaptive Training Tool for Critical Paper Reading. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.