AiModerator: A Co-Pilot for Hyper-Contextualization in Political Debate Video

Peter Andrews
MediaFutures, t2i Lab
University of Bergen
Bergen, Norway
peter.andrews@uib.no

Njål Borch
Schibsted
Oslo, Norway
njaal.borch@gmail.com

Morten Fjeld
MediaFutures, t2i Lab
University of Bergen
Bergen, Norway
morten.fjeld@uib.no

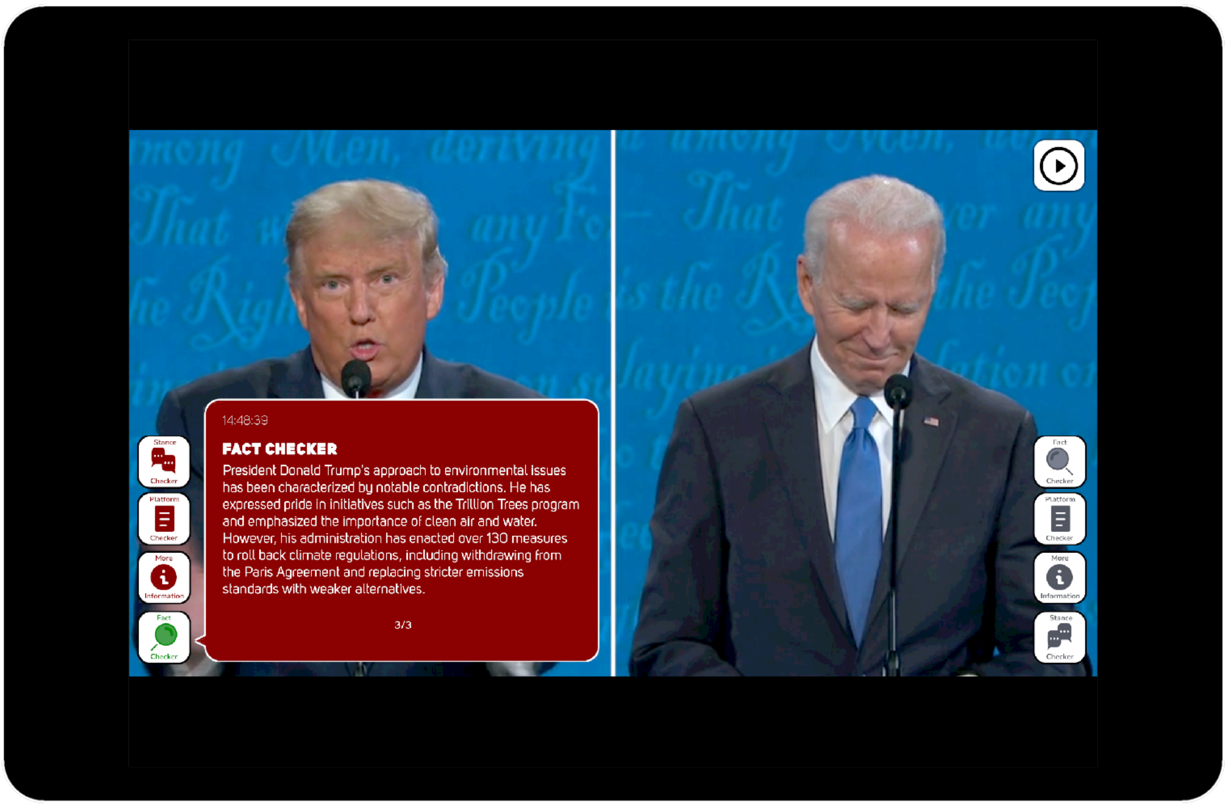

Figure 1: AiModerator is an event-driven co-pilot that delivers real-time, contextually relevant information to users. Its user interface overlays the video feed, enhancing accessibility. The functionality is categorized into primary and secondary features, designed to consolidate information and enhance user engagement. Primary features serve as the main interaction gateways and include the `Fact Checker`, `Stance Checker`, `Platform Checker`, and `More Information`. Secondary functions, such as `Explore Topic`, `Opinion Poll`, and `React`, enable continued interaction based on the primary features. AiModerator enhances user engagement and promotes a deeper understanding through hyper-contextualization of political debate video. Video and Image courtesy of CBS, © 2020. All rights reserved.

Abstract

Political debates are essential in political discourse for democratic societies. Advancements in technology have significantly transformed the structure of political debates, the ways in which politicians communicate, and the platforms through which audiences engage with them. Originally a forum for improving understanding, political debates have increasingly favored theatrics over substance,

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712148

risking young adult disengagement. To bring substance back to this medium we developed AiModerator, a political debate co-pilot acting as a Multimodal Conversational Agent (MCA). AiModerator aims to promote engagement while improving understanding by analyzing video content to provide contextually relevant information. This consolidated information facilitates understanding while keeping users synchronized with the debate viewing experience. Our system builds upon multimodal techniques, integrating computer vision and large language models to demonstrate ways of improving content delivery and engagement. AiModerator's backend system extracts events from identified speech data, allowing the user to interact with these events through a touch interface on an iPad application. We address three key topics: evaluating young adults' engagement, satisfaction, and preference compared to traditional second screening, and determining whether AiModerator can improve subjective understanding. To evaluate these measures we conducted a mixed-method evaluation (n=20) within-group design A-B study. Our analysis found AiModerator excelled in promoting engagement and satisfaction while delivering clear, contextually relevant information to the user which improved their understanding of debate topics more than the second screening mode. Our qualitative analysis offers broader insights, particularly in terms of a trade-off between automation and information consolidation versus autonomy and control.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; **Natural language interfaces**; *Usability testing*; **Human computer interaction (HCI)**.

Keywords

Multimodal Conversational Agents, Hyper-contextualization, User Experience, Natural Language Processing, Information Accessibility, Political Engagement, Interactive Video, Political Discourse

ACM Reference Format:
Peter Andrews, Njål Borch, and Morten Fjeld. 2025. AiModerator: A Co-Pilot for Hyper-Contextualization in Political Debate Video. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3708359.3712148

1 Introduction

Political debates are a crucial element of political discourse and consist of two main attributes. The first one aims to enhance the viewer's knowledge and understanding of policies, as well as the stances of political parties and candidates. The second one focuses on showcasing each candidate's valence traits—such as charisma, likeability, and general appeal—to attract viewer support. While both attributes serve their purpose, the latter has increasingly taken precedence, resulting in a loss of substantive information. In response, to add additional context, viewers often use a second screen to engage in political discourse via social media forums [29] and online resources. Those seeking additional fact-based information use second screening to obtain and share supporting material [31]. This information is then circulated on social networks to inform followers, express opinions [14], and solicit others' views [29]. The use of second screening during political debates has been shown to enhance viewer engagement by making the experience more interactive and informative. However, while second screening has helped improving political knowledge [60], the presentation of information can affect viewer cognitive load [67], possibly leading to information overload and distraction from the primary video stream [53]. Structuring complex information can mitigate these challenges and improve viewer comprehension. Research has shown that it also enhances educational value [1]. Retrieval-Augmented Generation (RAG)-based methods, which rely on Large Language Models (LLMs) to combine information retrieval with generative capabilities, actively structure large text corpora by extracting and summarizing key points.

To address the challenges and enhance comprehension of political debates, we introduce AiModerator, which employs a RAG architecture to structure information within a single device. AiModerator is a co-pilot for hyper-contextualization in political debate video. Drawing upon the term `moderator`, AiModerator mirrors the roles typically associated with a human moderator. Just as a human moderator oversees a debate by directing the discourse and clarifying points of contention, AiModerator performs a similar function through technological means. Instead of interjecting and guiding the debate like a traditional moderator, our novel invention extracts and synthesizes information from the debate. It then presents this information to the user in an accessible manner, enhancing both comprehension and engagement.

AiModerator is an event-driven system, which leverages Natural Language Processing (NLP) and LLMs to extract information from video stream. From the event data, AiModerator queries a database based on the user interaction to present clear accessible information. This information is hyper-contextualized, meaning it aligns precisely with the current discussion. Users can fact check statements in real time while simultaneously accessing additional information about party platforms, candidate stances, policies, people, and legislation. Moreover, AiModerator provides visual feedback through opinion polls and live reactions. In this manner, AiModerator functions as a Multimodal Conversational Agent (MCA).

AiModerator was tested on the second 2020 presidential debate between Donald Trump and Joe Biden during the Autumn of 2024. The debate was chosen due to its known concentration on rhetoric, inconsistent nature and lack of focus [65]. We tested the prototype by conducting a user study with young adults, a group that frequently disengages from politics [13]. Our study addresses the following research questions through conducting a mixed-method evaluation on data from a within-group design (A-B) study with twenty participants:

RQ1) How do users perceive the user experience of AiModerator for political debate viewing compared with second screening, particularly in terms of engagement and satisfaction?
RQ2) How can AiModerator help young adults improve their understanding of specific topics being discussed in a political debate video?
RQ3) Do users prefer AiModerator to the second screening mode, and if so, why?

Our contributions are threefold: 1) the creation of AiModerator, a political debate co-pilot MCA designed for hyper-contextualization;

2) the carrying out of a comprehensive study to evaluate the benefits and challenges of using MCAs while watching political debate videos; and 3) providing a novel and innovative solution to engage viewers actively in political discourse.

Our paper is structured as follows: first, we introduce political debates as a platform to improve understanding. We then outline a pre-design workshop, followed by the system design. Next, we present AiModerator's interactive features before reporting on the user study and its results. Lastly, we present our discussion, limitations, and conclusion.

2 Related Works

In this section we outline the purpose of political debates through the lens of improving understanding before considering interactive platforms built upon them. We then explore interactive video and its use cases in current research trends.

2.1 Political Debates - A Platform for Understanding

Political debates aim to improve public comprehension and assist voters in aligning with political movements [19]. Research indicates that the format of debates significantly affects information retention and knowledge. Druckman [24] found that participants who watched video debates retained more information than those who listened to audio-only formats, likely due to visual cues reinforcing the content. Research based on the American National Election Studies surverys (1976-2000) revealed debates increased issue knowledge and salience [8, 9]. Similarly, Zhu et al. [79] demonstrated that debates significantly impact understanding of candidates' issue stances, though not for undebated issues. Highlighting stances is particularly important for lesser-known candidates [10]. Evidence shows that political debate videos significantly affect issue knowledge, candidate preference based on issue stances, and identification of candidate positions [7].

Political debates also shape subjective measures, like Political Information Efficacy (PIE), which gauges confidence in political understanding and participation. A study of the 2004 Bush-Kerry debate showed an increase in PIE after the debate, although it dropped after the election [49]. McKinney & Rill [51] found that traditional and CNN-YouTube debate formats increased PIE.

However, recent debates have sometimes hindered understanding and engagement due to candidate behavior overshadowing substantive discussion. Rowland [65] critiqued the 2020 Trump-Biden debates for inconsistent behavior, lack of substantive policy discussions, and excessive rhetoric, all of which undermined informative value. Marin-Llado & Tornero [46] noted that media trends increasingly prioritize spectacle over political discussion. This in turn exacerbates societal polarization [37]. In the UK Gorkovenko & Taylor [32] found that second screening applications revealed a lack of substantive information and lack of trust.

2.2 Interactivity and Political Debate

While second screening can enhance accessibility to relevant information, it often distracts viewers with unrelated content, reducing focus on the debate. Research indicates that second screening during political debates is widespread. Anstead & O'Loughlin [4] observed increased viewer engagement on X during the 2009 debates, signaling the rise of media hybridity and active video engagement. Gorkovenko & Taylor [30] analyzed 38,569 tweets from the 2015 UK general election. They found that second screeners used X to share opinions, humor, and provoke others. Shah et al. [69] noted that visual elements had a more consistent impact on social media responses than tonal factors, highlighting the role of nonverbal cues in driving second-screen interactions. Motivations for using a second screen for social media include learning, gauging others' opinions, sharing information, voicing opinions, and boosting ego [29]. Chadwick [14] found that acquiring and sharing information was more important to second screeners than influencing others. McKinney et al. [50] reported that second screeners with higher pre-debate PIE tweeted more during the 2012 Republican primary debates. Ran & Yamamoto [60] discovered that task-relevant second screening improved factual knowledge recall, whereas task-irrelevant use was detrimental.

Other approaches aim to enhance engagement through collective means and collaborative analysis tools [12, 22, 33]. Deb8 [12] reintroduced context into debate videos by enabling collaborative information collection and argument chain creation; however the interface was not user-friendly enough for general audiences. Liddo et al. [22] engaged groups of viewers with flashcards, eliciting more emotional responses, while Gorkovenko et al. [33] used physical printers to connect participants' living rooms to networked groups. The tactile nature of printouts and pseudonymous interactions were found to encourage reflection and free sharing without online biases.

Innovations in interactive political debate platforms include integrating interactive video tools. Democratic Replay [57], provided argument maps, debate-rule compliance checks, interactive flashcards, and fact checking. A user study with 29 participants showed that users with lower levels of interest rated the system lower than more interested users [56]. A larger study with 113 participants found Democratic Replay improved seven out of nine sensemaking factors compared to standard broadcasts [23]. However, Democratic Replay's lack of real-time automation and its dependence on viewer interest limits accessibility.

2.3 Interactive Video

Research has shown interactive video's potential to boost motivation and improve learning outcomes without increasing cognitive load [44, 59, 70]. Shelton et al. [70] found that embedded quizzes significantly improved quiz scores and maintained viewer attention. Desai & Kulkarni [44] reported that 94.2% of participants achieved higher grades with interactive videos. Priyakanth et al. [59] observed improved learning and engagement in a MatLab course. However, the application of collaborative learning in this context remains underexplored. Kazanidis et al. [41] proposed using augmented reality to enrich learning experiences but did not conduct a user study to assess its impact.

In the entertainment industry, interactive video has expanded with innovations like Netflix's Bandersnatch, which offers nonlinear narratives responsive to user choices [63, 64]. Despite positive usability ratings, criticisms include limited autonomy due to binary choices [63]. Rezk & Haahr [62] suggested that modeling users' psychological states (invisible agency) could enhance narrative adaptability.

Interactive video also enriches storytelling by integrating additional data, thereby increasing user engagement and motivation [36]. Companion apps synchronize with primary videos to provide context and promote critical thinking [26, 71]. For example, Ma et al. provided additional context through 3D environments [47]. Cox et al.'s [21] CAKE application dynamically adapted a cooking programme in real-time to user interactions [21]. New modes of interaction have made an impact on interactive video. VR experiences like "Tell Me, Ingre" [39] and 360 video applications [15] demonstrate interactive video's potential to foster empathy and immersive navigation. In the sports domain, gaze-moderated interaction [17], voice interaction [45], and direct interaction with MCA commentators [3] have dynamically adapted the presentation of additional information through embedded visualizations.

Advances in LLMs have further transformed interactive videos by enabling the synthesis of relevant new information [42]. Ma et al. [47] created chatbots embodying characters that adapt stories to user emotions, while Jorgensen et al. [40] integrated additional information to provide more context to responses. Xu et al. [75] developed interactive science program chatbots for children, enhancing learning and meaningful interactions. Tanprasert et al. [73] found chatbots increased critical thinking but did not alter users pre-existing opinions.

3 Pre-Design Workshop

In the early design phases of AiModerator an HCI researcher collaborated with a political philosophy researcher to conduct a workshop to gather expert insights and address potential for innovation across a multi-disciplinary group. The two hour workshop was attended by four HCI researchers, one political scientist, a political philosopher, a political theorist, and two data scientists from a local broadcaster TV2. The workshop employed a semi-structured format with the HCI researcher and political philosopher overseeing discussions aimed at designing specific interactive features to aid understanding, critical thinking, interest, and engagement in political debate.

3.1 Workshop

The workshop began with a brief discussion to determine whether using Norwegian citizens as participants were suitable for watching a US presidential debate video. It was agreed that the nationality of the participants was not a topic, as the study aimed to improve general understanding of political topics. Moreover, unfamiliarity with the intricacies of US policies made non-Americans even more appropriate for the study. Further supporting the choice of participants, political scientists emphasized the global impact of US politics and the spectacle surrounding them.

The discussion then shifted to consider political debates and their stakeholders, identifying three key stakeholder perspectives:

- **Audience**: Seeking greater understanding, factual clarity, and entertainment.
- **Politicians**: Aiming to persuade and engage voters.
- **Broadcasters**: Focused on audience engagement and viewership metrics.

As our project's aim was to facilitate understanding, we discussed concepts of enhancing user experience, including personalized experiences. However, concerns were raised about the risks of personalization leading to fragmentation, filter bubbles, and echo chambers. The discussion also touched on current topics in U.S. and U.K. politics, noting how two-party systems contribute to societal polarization and how populism influences political debates. There was consensus that the recent use of rhetoric and personal attacks has created incoherent debates which confuse the populace, lead to disengaged citizens, and ultimately damage democracy.

The workshop also explored what features an interactive system could offer to better facilitate understanding. Several potential functions were considered:

- `Fact Checker`: Providing real-time verification of claims.
- `Highlight Dog Whistle`: Identifying coded language meant for specific groups.
- `Contrast Argument`: Presenting opposing viewpoints to a claim.
- `Exploring Topic`: Encouraging users to delve deeper into topics related to the discussion.
- `What-If Scenarios`: Allowing users to explore alternative outcomes.
- `Opinion Poll`: Showing users where their opinions lie within the populace.
- `Reaction Map`: Displaying live reactions from different regions.
- `Platform Checker`: Relating the current topic to wider political party viewpoints.

We speculated on these features, considering the need to balance user engagement with the provision of accurate information. The goal was to enhance understanding without leading to further fragmentation.

After the workshop, a political philosopher and an HCI researcher reviewed workshop notes to finalize the AiModerator's features. They selected debate video segments and evaluated each proposed feature's potential for meaningful interaction. The decision to omit the detection of `Dog Whistles` and the `What-If Scenarios` was due to their minimal presence and potential distraction, respectively. They refined the `Contrast Argument` into a `Stance Checker`, recognizing the direct opposition often found between Republican and Democratic leaders, which could clarify leadership stances on key issues. To address frequent use of acronyms and ambiguous terms, they added a `More Information` function for clarity. Considering the study's scope and the need to avoid overwhelming users, the `Reaction Map` was also removed. The final set of seven features - `Fact Checker`, `Stance Checker`, `Platform Checker`, `More Information`, `Explore Topic`, `Opinion Poll`, and `React` - were chosen for their ability to enrich the viewing experience, enhance understanding, and engage users within a limited timeframe, balancing informative content with interactive elements.

4 AiModerator System Design

The AiModerator system architecture comprises of both frontend and backend components. The backend is responsible for ingesting raw debate video data, while the frontend manages user interactions.

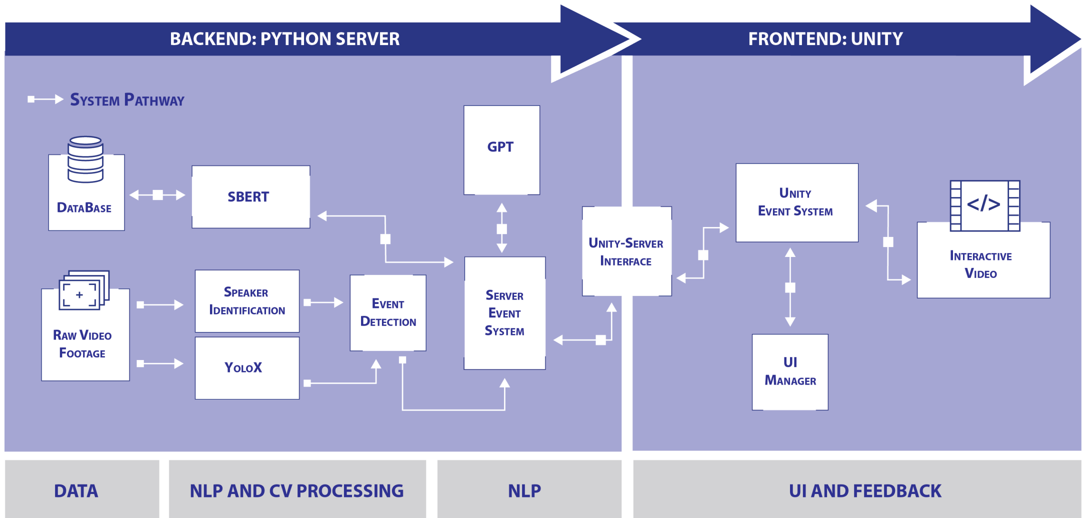

Figure 2: The AiModerator system processes the raw video feed using WhisperX [5] for transcription and speaker identification. Concurrently, YOLOX [27] tracks the two political candidates, and this information is fed into the event system. Events detected from the dialogue are then transmitted via TCP to the Unity game engine, which hosts the user interface designed for touch interaction. When a user interacts with the system, it retrieves relevant textual data from the database that provides additional context through the GPT module.

4.1 Backend System

The backend system was developed in Python, utilizing an object-oriented methodology. This architecture facilitates comprehensive handling of tasks - including NLP, Computer Vision (CV), event detection, API communication, and event management. This design ensures efficient processing and integration of these components within the system.

4.1.1 Database. We built the database using two primary resources: Wikipedia and the Serphouse API. To minimize API usage, we initially extracted events and keywords directly from the video content. For this project, we maintained a static database that contained all relevant information for the system. We saved Wikipedia articles under the keyword names used to search the API. Serphouse searches were restricted to data up to 20-10-2020, two days before the debate. We created feature embeddings for each article using SentenceBERT [61] encoder. We also encoded the events identified by the system and used cosine similarity to query the database. The article most closely matching the event was then augemented to the prompt. Finally, the database includes extracts on security and immigration issues, sourced from the Democrats and Republicans' 2020 party platforms.

4.1.2 Candidate Visual Tracking and Speaker Identification. Visual tracking of each candidate was necessary for augmenting visuals. First, we implemented YOLOX [27] to detect objects in the video footage. We specifically filtered these detections to identify only "Person" and "Tie" categories. To track these identified objects throughout the video, we employed DeepSORT [74]. This tracking algorithm embeds the detected regions to maintain the identities of each object throughout the video sequence. For speaker audio transcription and identification we used WhisperX [5], to differentiate between speakers during the debate.

4.1.3 Event Detection. To detect events, we first parsed the audio transcription using GPT-4. We initialized GPT-4 with a system context specifically designed to extract new events from the data. Periodically, as the audio transcription was fed into GPT-4, we prompted it to summarize the content while identifying topics, viewpoints, actions, counterarguments, and keywords. Events identified during this process were then encoded with feature embeddings using SentenceBERT. Events that met a certain similarity threshold were discarded, as they were considered too similar to previously identified events.

4.1.4 GPT. For real-time user interaction, we implemented four separate instances of GPT-4o, which was both the most up-to-date and cost-effective option available during our user studies. We configured each GPT-4o instance with system contexts tailored to the primary features specific tasks. Each instance received unique prompts that were activated as users interacted with the respective features. To address latency imperfections, calls to the GPT API were preprocessed using multi-threading as soon as an event was detected; ensuring timely updates to the user interface. Furthermore, each feature maintained a log of its interaction history, ensuring follow-up operations were within the same contextual space.

4.2 Frontend System

Following the same object-oriented methodology the frontend system was instead developed in Unity using C# to support touch interactions. We chose to deploy the system on an iPad for ease of interaction. This choice was made to align with the growing

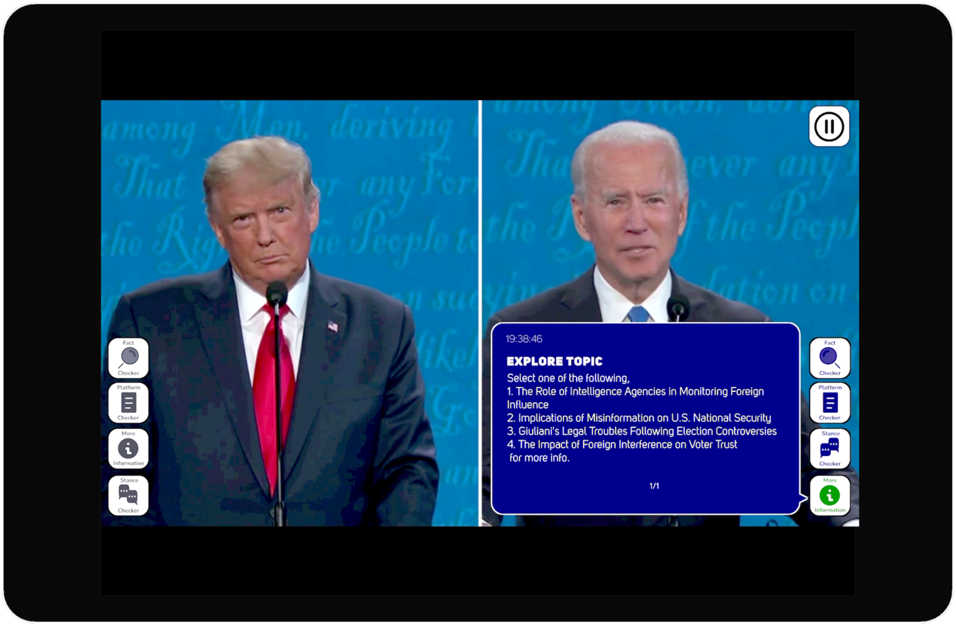

Figure 3: AiModerator: A Political Co-Pilot for Hyper-Contextualization in a Political Debate Video. Secondary features, such as `Explore Topic`, are activated through touch interactions with the active tab. Video and Image courtesy of CBS, © 2020. All rights reserved.

trend of watching video content on mobile devices, driven by their convenience and accessibility. We used Unity due to its robust tools for building interactive, touch-based applications and the team's familiarity with the platform and existing codebase. This choice enabled rapid prototyping and iteration, which was critical given the project's timeline. The app was published and pushed to an iPad using Unity and Xcode.

4.2.1 Unity-Server Interface. To facilitate communication between the Unity frontend and the Python backend, we utilized the NetMQ [20] library, which supports both C# and Python. TCP communication from the Python server to Unity employs a PUSH-PULL mechanism, which serializes and transmits images and textual responses generated by the GPT module. Conversely, TCP communication from Unity to the Python server utilizes a REQ-REP mechanism. This setup allows user interactions in Unity to trigger queries that have been sent to the Server Event System. The system then broadcasts these queries to the GPT module for processing.

4.2.2 Unity. Users could interact with primary (`Fact Checker`, `Platform Checker`, `Stance Checker`, and `More Information`) or secondary (`Explore Topic`, `Opinion Poll`, and `React`) features through a button and tab-based interface. Each tab operates as its own state machine, simplifying integration and maintenance by isolating functionality. All primary features (except `React`) reside in separate tabs, supporting gestures such as taps, pinches, and swipes for intuitive navigation.

4.3 Pilot Studies

We adopted a user-centered design for AiModerator, refining the prototype through iterative feedback from a pilot study with six Norwegian participants aged 20-28. Participants provided insights primarily related to the user interface, expressing difficulties in identifying which features were triggered, navigating secondary functions, and understanding their current depth in the tab system. In response, we labeled buttons, added visual prompts for secondary functions, and clarified text labels.

We also addressed language complexities in the prompts, particularly for lower summarized depth levels. One participants noted the lack of a chat history and the ability to return to previous tabs, we view this as a deliberate design choice to encourage users to stay current with the debate topics. Although future iterations of the system could explore the feasibility of incorporating a tab history feature, we deemed it unnecessary for the scope of this study.

4.4 Interaction with AiModerator

The AiModerator interface uses an MCA for interaction. As mentioned previously, GPT4o generates live feedback in response to user interaction. The user can interact with the system, triggering primary features by clicking on the corresponding button. Secondary features are triggered by touch interactions with the tab. Andrews et al. [3] found that users struggled to formulate questions to the MCAs when watching video content. Therefore, AiModerator automates this process by giving a selection of possible interactions. Our aim was to lower mental effort, assisting the user in retaining their attention on the video. When a user interacts with the system, the Unity interface sends the information to the Python server, which in turn facilitates interaction with the GPT API. As the system awaits feedback, the touch interface locks, and a waiting symbol that informs the user that a response is being curated appears. We also implemented a `depth-text` feature, which allows users to control the level of detail of responses provided by the MCA. The purpose of this feature is to enable users to select the depth of detail that best suits their comprehension needs. For our prototype we limited the available depth options to three levels: the first level provides a brief summary in a single sentence, the second offers a explanation of a few sentences, and the third delivers an in-depth summary in a short paragraph. We now outline each of the interactive features to explain their functionality and relevance.

4.4.1 Fact Checker. The `Fact Checker` was our most direct approach to addressing the lack of substantive information; it adds objectivity amid the uncertainty of various candidate claims. Fact checking can help promote trust in politics [32]. We designed the `Fact Checker` to explicitly communicate a verdict about whether the specific claims are accurate while including evidence and explanations.

4.4.2 Platform Checker. Our literature review revealed that recent political debates often prioritize theatrics over substantive political discussions [65]. As a result, debates become more focused on individual personalities rather than the parties and their policies. The `Platform Checker` aims to place debate topics within the broader context of political parties by retrieving information directly from each party's platform.

4.4.3 Stance Checker. While considering the broader party perspectives, it is still relevant to consider the individual candidates. Identifying stances on specific topics can be challenging, especially in debates where theatrics overshadow content. Therefore, the `Stance Checker` considers context from the event system and the article database to formulate and contrast the stances of both politicians.

4.4.4 More Information. Video segments chosen from the 2020 presidential debate contained ambiguous terminology and references to events, places, and names that may be unclear or unknown to many viewers. The `More Information` feature lists detected keywords and provides relevant information from the Wikipedia gallery.

4.4.5 Explore Topic. The first of our secondary features can be accessed by double-tapping anywhere within the tab. This allows continued interaction with the MCA. The MCA delivers a list of contextually relevant topics based on conversation logs. The user can then select one of these topics to receive related information while staying within the current context. As the debate progresses both the context and the topics update to reflect the current discussion.

4.4.6 Opinion Poll. Partially inspired by the works of Gorkovenko et al. [29], we wanted participants to reflect on their experiences and share them with a wider community. The `Opinion Poll` presented users with a short question or statement and a selection of opinions that reflected either a general Republican or Democratic stance. When an opinion is selected, an animated star transitions to the debate candidate who best reflects that viewpoint. Afterwards, a bar chart which displays simulated responses for each of the selections appears in the tab.

4.4.7 React. Finally, the `React` feature allows users to express their feelings by posting an emoji in response to what a candidate has just said. This feature can be activated from any of the tabs by swiping up. When selected, the user chooses between a red emoji for Republican or a blue emoji for Democrat. The selected emoji then follows a randomized spline curve defined by tracking data.

5 User Study

We conducted a mixed-method user study (n=20) to assess young adults' experience with AiModerator. The study followed a within-subjects design (A-B) methodology. The two conditions studied were the two alternative modes of interaction: second screening and AiModerator. All participants worked in both conditions. The order in which participants undertook the conditions depended on the group to which they were assigned. Similarly, the video content (national security and race and immigration) was counterbalanced to mitigate content bias.

5.1 Participants

Twenty participants (6 males [30%], 14 females [70%]) between 20 and 28 years of age (M = 22.9, SD = 2.4) participated in the user study. The participants were from Norway (17 participants, 85%), Ukraine (2 participants, 10%), and the United States (1 participant, 5%). Political researchers in the pre-design workshop considered the inclusion of non-US native participants unproblematic, as the study focuses on enhancing general understanding of debate topics and user experiences of such tools. Non-US citizens were particularly suitable due to their reduced familiarity with the intricacies of US policy and the growing international awareness, impact, and interest in US politics. The participants were recruited through various means, including project promotions on multiple university course websites and in-person pitches delivered during university lectures. Participants were offered a gift-card suitable for local stores worth 200NOK as an incentive.

5.2 Setup

The study was organized into four stages: introduction and consent, second screening mode, AiModerator mode, and post-study interview. While the first and last stages were the same for all participants, the order in which users experienced the second and third stages depended on the group to which the participant was randomly assigned. Each participant was exposed to two video segments: one on national security and one on immigration and race. These topics were chosen due to their current international relevance.

5.2.1 Introduction and Consent. Participants were greeted and introduced to the structure of the study. They were verbally informed of their rights, the types of data to be collected, and how the data would be collected, processed and stored. Participants then read and signed the consent form.

5.2.2 Second Screening Mode. The researcher provided participants with a guided walkthrough on how to work within the second screening condition. Participants were instructed to use only the URL search bar in Safari to conduct searches. This limitation was necessary because a background program appended the date parameter "before:2020-10-20" to all searches, restricting results to those published before October 20, 2020—two days prior to the second presidential debate. Participants were allowed to access any link from the Google search results but were not permitted to follow links to any other pages. From the accessed page, they could either return to the search results or initiate a new search using the URL bar. Participants were informed that accessing Wikipedia pages from the Google results was acceptable. This exception was made because Wikipedia content could not be controlled due to its frequent updates. Moreover, the AiModerator condition included data sourced from Wikipedia.

After confirming that the participants understood the instructions, we asked them to conduct a test search while watching a six-minute video segment from the same debate on environmental issues. The researcher observed to ensure the participant adhered to the rules. Subsequently, the participants watched a new twelve-and-a-half-minute video segment featuring different footage. They were instructed that they could pause the video for a total of only three minutes. This limitation was imposed to ensure the study could be completed within the allotted time (1 hour 30 mins). After viewing the video, the participants completed a Post-Viewing Questionnaire (see Section 5.3.2) and answered the Rating Scale for Mental Effort (RSME) question [80].

5.2.3 AiModerator Mode. The researcher conducted a guided walkthrough of the AiModerator mode, during which the participant completed the Interactive Features Questionnaire (see Section 5.3.1). After completing the walkthrough, the participants were informed of the three-minute pause limit before watching another twelve-and-a-half-minute video segment featuring new, unseen footage. They were encouraged to interact freely and as frequently as they desired. Afterwards they completed a Post-Viewing Questionnaire (see Section 5.3.2), the System Usability Scale (SUS) questionnaire, and the RSME question [80].

5.2.4 Post-study Interview. Before the interview, we asked each participant to state their preference by circling the corresponding mode on a form. The researcher then followed a semi-structured interview, questioning the participant about their experiences with each viewing method, their likes, dislikes, and reasoning behind their preference. The interviews were audio-recorded and lasted between 10 minutes and 18 minutes and 20 seconds.

5.3 Questionnaires

Participants completed an Interactive Features Questionnaire after testing each interactive feature. Moreover they completed a Post-Viewing Questionnaire after each mode of interaction, where all questions were designed to generalize to both modes of interaction. The questionnaires were inspired by Chen et al. [17] and Lin et al. [45], and adapted from Andrews et al. [3]. Due to changes in content and research aims, the questions were modified to better suit content accessibility, context, and clarity (Appendix A and Appendix B). Both questionnaires used a 7-point Likert scale to enhance sensitivity and detect subtle variations in user perceptions, enabling more precise and reliable feedback [43, 58].

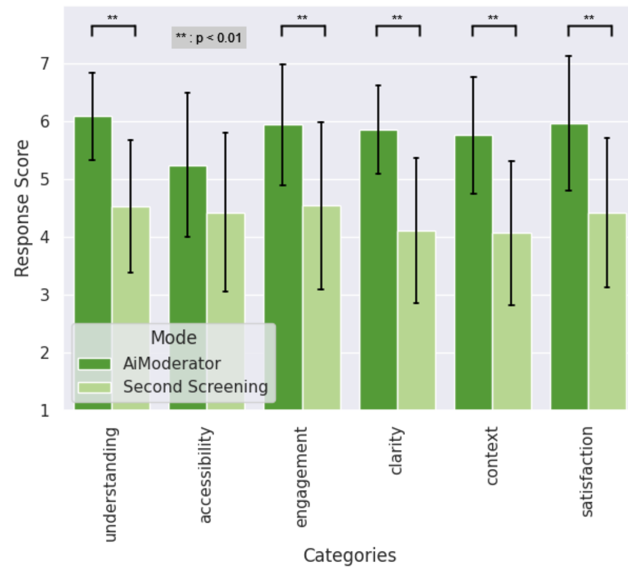

Figure 4: Post-Viewing Questionnaire Results. The results show a significant difference between the two modes, AiModerator and second screening, in the categories of "Understanding," "Engagement," "Clarity," "Context," and "Satisfaction," with AiModerator scoring significantly higher in these categories.

5.3.1 Interactive Features Questionnaire. This questionnaire (Appendix A) was designed to provide further insights into RQ2. The aim was to better understand which features of the co-pilot contributed to the perceived user experience of AiModerator. The questions for the primary features and the secondary feature, `Explore Topic`, consisted of six questions, while the remaining secondary features each consisted of seven questions. The first question for all features aimed to learn whether the user perceived the feature as relevant to debate viewing. If the score was low, it might indicate a reason for further lower scores. The other statements examined individual factors such as engagement, clarity of information, context, satisfaction, and future use. For the `Opinion Poll` and `React` features, questions regarding clarity and context were omitted and replaced with questions focusing on expression and reflection on viewpoints.

5.3.2 Post-Viewing Questionnaire. The questionnaire (Appendix B) was divided into categories. "Engagement" and "Satisfaction" were formulated to answer RQ1. "Understanding", "Context", and "Clarity" aimed to address RQ2.

5.4 Analysis

Our analysis was based on the Post-Viewing Questionnaire and RSME scale (both modes), the Interactive Features Questionnaire and SUS (AiModerator mode only), and overall preference. Audio recordings were taken during the post-study interviews and transcribed using Whisper STT before being checked and refined by a researcher. Qualitative analysis was performed using thematic analysis [11] to uncover recurring themes. For the quantitative analysis, we took the mean scores of questions contributing to each category: "Understanding", "Accessibility", "Engagement", "Clarity", "Context", and "Satisfaction." We tested each category for normality before conducting paired t-tests between the two modes.

Table 1: Total Preference for Each Mode of Interaction.

Mode	Total Preference
AiModerator	18
Second Screening	2

6 Results

This section presents the quantitative and qualitative results of our user study.

6.1 Quantitative Results

Figure 4 reveals the results of the Post-Viewing Questionnaire. Before statistical analysis, a Shapiro-Wilk test ($p > .05$) and Q-Q plots confirmed normality across all categories, indicating the data was suitable for parametric T-test. We conducted paired T-tests for each category of the Post-Viewing Questionnaire between the two modes. We found that the scores were significantly higher with the AiModerator than second screening in "Understanding" ($df(19), t = 5.22, p < 0.001$, Cohen's $d = 1.168$), "Engagement" ($df(19), t = 3.98, p = 0.001$, Cohen's $d = 0.88$), "Clarity" ($df(19), t = 4.75, p < 0.001$, Cohen's $d = 1.06$), "Context" ($df(19), t = 4.78, p < 0.001$, Cohen's $d = 1.07$), and "Satisfaction" ($df(19), t = 4.3, p < 0.001$, Cohen's $d = 0.96$). "Accessibility" ($df(19), t = 2.07, p = 0.057$, Cohen's $d = 0.45$) was at the statistical tendency level. The mean SUS score for the AiModerator was 77.5 (*"GOOD"* in adjective rating).

Table 1 displays the total preference count for AiModerator and second screening modes. A chi-squared test indicated a significant difference in preference between AiModerator and second screening ($\chi^2(1, N = 20) = 20.00$, $p < .001$). These results suggest a strong preference for AiModerator over second screening.

We conducted a Shapiro-Wilk test, which confirmed that the RSME data were normally distributed. A subsequent paired samples t-test ($df(19), t = -0.319, p = 0.753$, Cohen's $d = -0.069$) revealed no statistically significant difference between the RSME ratings for the two modes. Figure 5 shows the median RSME for the AiModerator is slightly higher than for the second screening, indicating participants reported marginally higher mental effort when using the AiModerator.

Figure 6 shows the likert responses for the Interactive Features Questionnaire. Overall results are extremely positive. All primary features and the secondary feature display ratings of slightly agree to strongly agree. The secondary features designed for active participation display more variance in results. We can observe `Opinion Poll` outperforming the `React` feature, which was the lowest rated of all features. Considering clarity and context, participants mostly felt the information was clear and relevant, and enhanced understanding of debate topics ($C3_n \geq 95\%, C4_n \geq 95\%$).

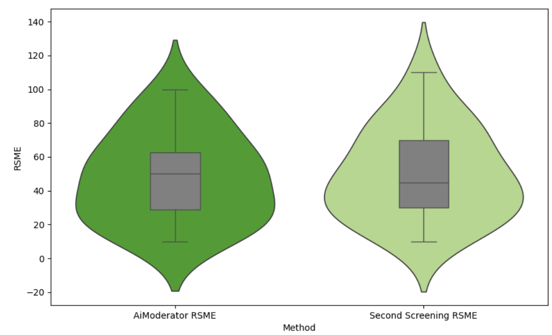

Figure 5: Rating Scale Mental Effort (RSME) Violin Plots for AiModerator and Second Screening. The box plots within the violins show the median response, with AiModerator having a higher median than second screening. The interquartile range (IQR), represented by the box, indicates a similar spread in both modes. The violin shapes represent the kernel density estimation, showing the data distribution, with both modes displaying similar distribution tendencies.

6.2 Qualitative Results

We now present our qualitative findings from the post-study interviews.

6.2.1 Usability and Understanding. Participants appreciated the freedom of second screening to explore aspects of interest (P08). However, many found navigating unstructured online resources challenging and increasing mental effort due to difficulties in finding relevant information and assessing source credibility (P07, P12, P16). P15 and P16 noted that identifying politicians' stances was problematic, suggesting second screening may not effectively aid understanding without significant effort. In contrast, the AiModerator system's timely delivery of consolidated, context-relevant information reduced effort and enhanced user experience. Participants felt it made information easier to digest and increased engagement with the debate content (P06, P15). P06 highlighted that "the accessibility ... lowered the barrier for actually engaging and trying to understand what was going on." The ability to control the depth of information balanced conciseness with informativeness, with depth level preferences varying among participants (P05, P14). Automation of content delivery further enhanced engagement, helping participants focus more on the debate (P06). While many appreciated the automation, some expressed frustrations with its delivery. P17 found the information surface-level and lacking depth, and P06 compared it to "fast food" —quick but lacking substance.

6.2.2 Engagement and Interactivity. AiModerator consolidated information into a single platform, enhancing focus on the video content (P05, P09). P09 noted, "It had a UI on top of the video ... it helps you focus more on [the] video." This unification made it easier to stay synchronized with the debate. Contextually relevant content further aided engagement, helping participants comprehend without lagging behind (P01, P14, P15). The system's prompts

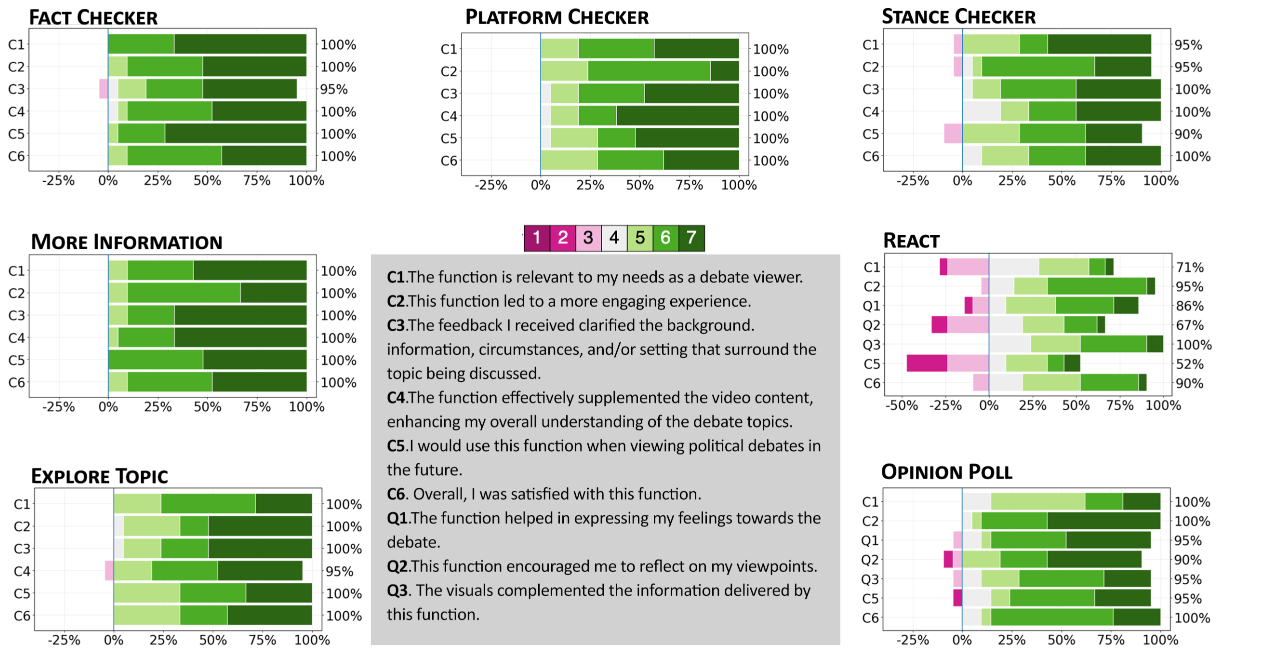

Figure 6: Interactive Features Questionnaire Rating Results for each Interactive Feature in AiModerator. The visualization follows methodology from [3, 17, 45].

encouraged exploration of topics participants might not have considered (P02, P09), though some found them distracting. P20 felt "sometimes like I wanted to interact with it, even though I didn't actually feel a need to." Interactive features like `Opinion Poll` and `React` enhanced engagement for some, providing a sense of involvement (P09, P15). P09 expressed "It's just a lot more fun ... I feel like I get to be a little part of it." However, not all experiences were positive; some felt the polls were too basic or caused discomfort when their views differed from the majority (P02, P11). For these participants, fear of being an outlier influenced their willingness to take part. P02 felt uncertain when his/her views differed from the majority:

> "Also maybe second thinking myself ... like wanting to agree with the masses instead of having my own opinion."

This conflict led to stress and disengagement from the `Opinion Poll` feature, distracting from the political debate.

6.2.3 Clarity. The `Fact Checker` was highly valued for verifying facts in real-time, allowing participants to stay informed about candidates' truthfulness (P01, P10). P01 noted it enabled them to "quickly either confirm or deny [statements] and then continue engaging with the debate." However, some felt it lacked depth or did not always activate when needed (P16, P17). P17 expressed it was "a bit too surface-level and lackluster at times."

The `More Information` feature provided context to complex topics, assisting participants in clarifying ambiguities (P15). P15 noted:

> "It helps clarify when they bring up situations because they only say the name ... they don't say what happened fully."

While the information sometimes piqued interest for further research, it wasn't always comprehensive enough to form a full opinion (P13).

AiModerator enabled participants to prioritize substance over rhetoric, focusing on context-relevant information (P01, P07). However, some found it distracting them from noting non-verbal cues essential to understanding the debate (P06). Others appreciated shifting focus away from distracting visual elements (P19, P02).

Participants noted past struggles identifying stances from political debate video alone (P04, P05, P10, P14, P16); others - P11, P13, and P18 -found it less difficult.

> "[The Stance Checker] focused information ... without the confusion that can arise from reading articles that try to summarize everything." (P19)

Similar to the the `Stance Checker`, the `Platform Checker` also improved the accessibility of specific information. P09 found that this feature helped in identifying the current topic within a wider scope, thus helping to understand party alignment.

6.2.4 Multitasking. Second screening increased mental effort and reduced focus on the debate as participants struggled to access reliable information in real-time (P03, P05). Some found multitasking with AiModerator still challenging (P03, P05); others felt it enhanced focus due to the unified platform (P01, P09, P15). Participants managed multitasking with AiModerator by controlling information intake through depth-text features (P05, P08), skim-reading (P12), selectively interacting with features (P07, P10), or engaging only when interested (P19). Personal habits influenced participants' ability to multitask effectively, with some finding it straightforward due to accustomed behaviors (P14); while others did not feel burdened by multitasking demands (P06).

6.2.5 Autonomy and Control. Second screening offered enhanced autonomy and control, allowing participants to search for specific topics and formulate their own questions (P01, P12). However, this came with increased mental effort since actively pursuing information was time-consuming and mentally taxing (P12). In contrast, the interactive system restricted options to predefined content, limiting the ability to find specific information (P03, P16). Thus, while second screening provided greater autonomy, it required extra time and mental effort (P07) navigating sources and could lead to getting lost in choice (P03).

6.2.6 Trust and Transparency. Participants valued knowing their information sources during second screening (P01). They desired more transparency in AiModerator, wanting references to verify information quality (P06, P09). This lack of clear sourcing led to uncertainty (P10). P20 noted trust depended on the platform's neutrality, highlighting concerns that media biases could harm the democratic process if not addressed.

7 Discussion

Our discussion is formulated around our three research questions, which we will approach individually.

7.1 AiModerator's User Experience

We now consider the user experience of AiModerator within the scope of RQ1. Our quantitative analysis revealed significant effects across user experience categories, including engagement and satisfaction. However, the accessibility attribute was at the statistical tendency level and therefore not considered statistically significant. Our qualitative analysis revealed participants found the system "quite intuitive to use" (P14), reflecting the SUS rating of 77.5.

7.1.1 Synchronization facilitating Engagement. AiModerator facilitated engagement through synchronization and the unification of content accessibility into a single platform. Based on our analysis, we have identified two variations of synchronization:

- **Video Synchronization:** Retaining focus on the video itself.
- **Context Synchronization:** Accessing topically relevant information driven by the current discussion.

Synchronization and unification of content within AiModerator enhanced user engagement by streamlining focus on the debate. While this approach generally supported multitasking, variations in individual abilities affected how participants managed additional information streams. Feedback regarding contextually relevant information was overwhelmingly positive both in the quanititative and qualititative data. The content was clear and relevant to many participants, indicating content synchronization enhances engagement.

However, it is crucial to balance the depth and breadth of additional content to maintain synchronization with the video. Previous research has documented that interactive platforms can distract from video content [25, 28, 35, 56]. Our findings suggest that whether an interactive system detracts from the video depends on participants' multitasking abilities and strategies for managing video multitasking, indicating a more complex relationship than initially anticipated.

Future iterations should focus on optimizing the granularity of information to better match user needs and reduce mental effort. Incorporating adaptive algorithms that adjust the depth of information based on user interaction patterns and video discussion pace could enhance personalization and efficacy. For example, user profiles could be dynamically constructed and utilized for in-prompt augmentation or adapted using the fusion-in-decoder model, as demonstrated in the LaMP framework [66]. Such refinements would promote engagement, making it more intuitive and less disruptive, especially during complex multitasking scenarios.

7.1.2 Engagement and Satisfaction. AiModerator's interactive features contribute an additional layer of engagement which affects satisfaction level. Despite the lower ratings for the `React` and `Explore Topic` features, their potential impact on engagement should not be dismissed. The `React` feature, though not widely embraced (Figure 6 Q3 $\geq$ 52%), offers immediacy and emotional connection by enabling real-time reactions. However, its limited use suggests misalignment with user preferences or perceived irrelevance, potentially due to concerns about distractions or limited impact in a lab setting.

Given Shah et al.'s [69] findings that candidates body language significantly influence social media responses, it is possible that AiModerator diverted attention away from the candidates, thereby reducing the emotional responses of participants. Such a shift in focus suggests that while AiModerator can enhance the informational aspect of the debate, it may do so at the cost of emotional engagement.

The `Opinion Poll`, despite being among the lower-rated features, showed a high satisfaction rate (Figure 6 Q3 $\geq$ 95%). Participants appreciated how it prompted them to reflect on their own opinions and easily identify which candidate their views aligned with. However, some participants felt uncomfortable when their views diverged from the consensus, risking disengagement. Emphasizing diverse perspectives or highlighting minority views could help mitigate this.

These observations indicate there may be a trade-off regarding emotional engagement vs objective viewing. Users may feel less emotionally connected to the candidates or the debate topics because they are viewing the debate from a more objective standpoint. By providing objective analysis through features such as the `Fact Checker`, AiModerator encourages a more analytical perspective. While this enhances understanding, it can also distance users from the emotional elements that promote engagement and active participation. To address this, future designs should place greater emphasis on supporting the emotional aspects of the debate. Customizable interaction preferences and personalized expressions can enhance emotional engagement, while a supportive online community, facilitated through real-time discussions, can further enrich emotional responses and satisfaction.

7.2 AiModerator's Impact on Understanding

We now consider AiModerator through the lens of RQ2. Our quantitative analysis revealed large significant effects in understanding, clarity, and context, with AiModerator scoring higher than second screening. We attribute this result to AiModerator's various features, which consolidate information into an easy-to-interpret, structured format. For example, the delivery of information in a structured manner with features such as the `Stance Checker` and `Platform Checker`, builds familiarity over time, further aiding understanding. Alshaikh et al. [1] found that their LLM-based tool, which incorporates structured information through various modules, had a positive impact on learning, engagement and content clarity. These findings support our observation that structured information can facilitate understanding. Moreover, our results support the notion that unstructured information from second screening can detract from the primary content [53]. AiModerator's structured approach appears to mitigate this issue.

These findings suggest that integrating structured informational tools into live broadcasts can improve viewer engagement and understanding in political debates. However, some participants felt the information was sometimes superficial, indicating a need for deeper content. More advanced structuring techniques, such as visualizations, graphs, and tactile prompting algorithms (which refine LLMS outputs) [38], can produce clearer, more coherent responses. Visual tools like mind maps or graphs also help users quickly grasp complex information, enhancing their engagement with the debate.

Our findings indicate that participants found the `Fact Checker` and `More Information` features especially useful for clarifying ambiguities, addressing misconceptions, and decoding acronyms. Our evidence supports Nyhan [54], suggesting that fact checkers can improve factual belief perception by reducing misperceptions. However, York et al. [76] found that fact checking can overwhelm users or highlight complexity, causing confusion rather than clarity. Realizing how difficult it is to discern truth may erode users' confidence in their political understanding, leading some to disengage from tools like AiModerator. An interesting synergy occurs when considering the `Opinion Poll`. The collaborative nature of this feature caused self-doubt in some participants when their opinions did not align with the consensus. Consequently, perceiving a viewpoint as more correct can negatively impact one's judgment and confidence. Therefore, interactive features like the `Fact Checker` and `Opinion Poll` must be designed in a way that supports users' confidence without overwhelming them. For instance, presenting fact checking information with clear explanations and actionable insights—rather than merely highlighting discrepancies—could help maintain engagement and trust.

7.3 Preference

Considering RQ3, there is a clear preference for AiModerator, highlighting its success in meeting user needs for engagement and understanding. However, the desire for autonomy and transparency expressed by some participants indicates a need for balancing automation and information consolidation with autonomy and control. The absence of source attribution in the co-pilot system made some participants uneasy about the credibility of the information provided. While the system delivered relevant information tied to points in the debate, it limited users to predefined content without revealing its origins. This lack of transparency led to frustration and skepticism, as users could not confirm whether the information was reliable.

To address these limitations, we recommend enhancing transparency to alleviate uncertainties regarding credibility. Integrating source citations directly into the content delivery process would allow users to assess the reliability of the information. Moreover, incorporating the concept of variable autonomy into AiModerator could help balance automation with user control, addressing participants' desires for both ease of use and autonomy. Variable autonomy refers to systems where the level of autonomy can be adjusted dynamically based on context and user preferences [52]. By allowing users to adjust the level of automation, AiModerator could provide a more personalized experience. For instance, AiModerator could offer different modes of interaction:

- **Automated Mode:** The system proactively delivers contextually relevant information with minimal user input.
- **Manual Mode:** Users direct the system by formulating specific queries or selecting topics of interest.
- **Hybrid Mode:** A combination of both, where the system suggests information but allows easy user customization.

This adjustable autonomy empowers users to tailor the system's operation to their needs and preferences, enhancing user satisfaction.

Similarly, variable transparency can address trust concerns by letting users adjust how much of AiModerator's operations they see, such as source attributions, reasoning processes, and confidence levels, to verify credibility. This aligns with Heer's [25] principle of balancing automation with human agency, where AI offers suggestions while the users retain control. By presenting curated information with embedded source links, AiModerator could promote user-driven exploration with automated assistance, preserving autonomy.

7.4 Accuracy and Trust

While this research focuses on the AiModerators UX, we recognize trust's impact on UX and that it is built upon truth, accuracy, and reliability. A rigorous evaluation of algorithmic accuracy is beyond the scope of this study. With tools like fact-checkers becoming increasingly accessible [18, 68], we focus on how users perceive and experience these tools rather than the underlying algorithms. Instead, we provide manual verification of AiModerator's fact check outputs [1], and a full interaction output log is available online [2].

The known hallucinations and biases of LLMs are concerning and continue to undermine user trust. Biases in particular occur from training data containing societal preconceptions. As recent studies demonstrate, LLMs can manifest gender, cultural, and political biases [6, 34, 77], with ChatGPT, for example, displaying pro-environmentalist and left-libertarian leanings [34]. Researchers must implement mitigation strategies to validate statements from LLMs to correct bias and prevent hallucinations. Various mitigation techniques exist, including prompt engineering [78], fine-tuning

[1] Refer to this link for manually verified fact-checker claims: https://anonymous.4open.science/r/IUI2025AiModerator-829C/1ClaimValidation.pdf
[2] Refer to this link to manually for an interaction output log :https://anonymous.4open.science/r/IUI2025AiModerator-829C/2InteractionLog.txt

[55], RAG [16], and error-correction mechanisms [2]. Each technique aims to prevent inaccurate or biased outputs or validate them post-generation. Manakul et al. [48] investigated consistency by comparing outputs generated through sampling multiple responses from GPT. However, the method worked at the sentence level, potentially missing finer details in sentences containing factual and non-factual information. Our approach employed a RAG-based framework to reduce hallucinations by prioritizing external content over GPT's internal knowledge. Chen et al. [16] demonstrated that the RAG technique mitigates hallucinations and enhances factual accuracy. However, RAG inherently shifts any biases or inaccuracies onto the external data sources. Therefore, its effectiveness is contingent on the quality of the data and the reliability of the encoder-based retrieval method. In designing our study, we intentionally withheld source accreditation to prevent participants from being distracted by evaluating the validity of the sources, allowing them to focus on interacting with the system. Nonetheless, trust and accuracy remain pivotal for LLM deployments in democratic societies, especially for sensitive domains like political discourse. Real-world applications increasingly reflect these concerns. For instance, TV2 recently implemented a conversational avatar with a downstream fact validator during the U.S. elections [72], illustrating LLMs usage in high-stakes broadcasting environments. We recommend future iterations of tools such as AiModerator to consider prevention techniques to promote trust and accuracy through hallucination control and bias mitigation. Further empirical research is essential to ensure these tools operate reliably and transparently.

8 Limitations

This study's main limitation is its lab-based setting, as users typically watch political debates in natural environments like their living rooms. Testing the prototype with only twelve and a half minutes of debate also limited our ability to observe long-term user interactions. Some participants reported fatigue from engaging with every feature due to novelty effects. Conducting longitudinal studies would allow us to assess sustained interactions and better understand AiModerator's long-term impact on objective understanding.

This study primarily focused on analyzing subjective understanding. While objective measures are generally more conclusive, it is important to recognize that politics inherently involves subjective interpretation.

Although the focus of this paper is not on the accuracy of AiModerator, we manually verified a subset of the system's outputs to ensure reliability. Future iterations of AiModerator will prioritize validation strategies (outlined in Section 7.4) and focus on optimizing system accuracy.

The participant base consisted of an international audience, primarily Norwegian. Political theorists and scientists validated this user base as appropriate for addressing the research questions posed in this study. However, future studies should consider the emotional involvement of participants in their native political systems, as this may influence interaction and engagement. AiModerator's system design is scalable and can generalize to the political systems of other countries. Nevertheless, multi-party debates would require re-evaluating the user interface to accommodate more politicians and complex video compositions.

9 Conclusions and Future Work

Our study compared and evaluated AiModerator, a co-pilot for political debate viewing designed to enhance understanding and engagement among young adults, with the popular second screening mode. This research establishes a benchmark in interactive political debate platforms and provides a foundation for future exploration of co-pilots in facilitating political discourse. AiModerator employs state-of-the-art CV and NLP technologies, allowing users to interact via a touch interface on an iPad application. This innovative approach reimagines how users engage with political debate video content.

We evaluated AiModerator using a mixed-methods, within-group (A-B) design, revealing a strong preference for it over second screening. Our findings demonstrate that AiModerator significantly improved subjective understanding and provided a highly engaging experience compared to second screening. The system achieved a SUS rating of 77.5, validating its usability. Additionally, AiModerator's interactive features, conceptualized during a multidisciplinary pre-design workshop, received high user experience scores, confirming their potential for effective use in real-time political debates.

While most participants preferred AiModerator over second screening, our analysis identified some key issues. In particular, results indicate a trade-off between automation and information consolidation versus autonomy and control. We suggest that developers consider variable autonomy in co-pilot systems to adapt to users' needs and preferences. AiModerator transitions the user into an objective viewing state, but developers should proceed with caution. Increased objectivity can reduce emotional engagement, which is paramount for active political discourse. Additionaly, fact checking tools and opinion polls may result in doubt in one's ability to discern objective truth, contributing to political disillusionment. Therefore, interactivity should be designed to enhance understanding without unintentionally reducing users' confidence in their opinions. AiModerator also risks diminishing emotional engagement by distracting from crucial nonverbal cues. Integrating advanced video analysis to highlight these cues and incorporating social networking features can help maintain emotional resonance.

Our future research will optimize the balance between these attributes to improve user experience. We will consider personalization and enhanced collaborative viewing methods while being cautious of potential issues related to polarization, such as echo chambers and filter bubbles. We aim to conduct longitudinal studies with such prototypes to measure the long-term impacts of hyper-contextualized media on user understanding and political involvement. Developers and designers can reference our findings and methodology for academic projects, news broadcasting, and political analysis.

Acknowledgments

This work was supported in part by the Norwegian Research Council (MediaFutures, 309339). It was also supported in part by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS), funded by the Marianne and

Marcus Wallenberg Foundation. We thank Philippa Beckman and Barbara Stuckey for proofreading and editing, and Izabela Krejtz and Jeng Jia-Hua for user-study design and data analysis advice. We also acknowledge TV2 for their collaboration in the pre-design workshop and continued dialogue regarding improving user experience with Human-AI Interaction. Finally, we extend our gratitude to Christopher Senf for his organization of the pre-design workshop, inspiring political discussions, and support on the AiModerator project.

References

[1] Rana AlShaikh, Norah Al-Malki, and Maida Almasre. 2024. The implementation of the cognitive theory of multimedia learning in the design and evaluation of an AI educational video assistant utilizing large language models. *Heliyon* 10, 3 (Feb. 2024), e25361. https://doi.org/10.1016/j.heliyon.2024.e25361

[2] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Learning From Mistakes Makes LLM Better Reasoner. https://doi.org/10.48550/arXiv.2310.20689

[3] Peter Andrews, Oda Elise Nordberg, Stephanie Zubicueta Portales, Njål Borch, Frode Guribye, Kazuyuki Fujita, and Morten Fjeld. 2024. AiCommentator: A Multimodal Conversational Agent for Embedded Visualization in Football Viewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. Association for Computing Machinery, Greenville, SC, US, 14–34. https://doi.org/10.1145/3640543.3645197

[4] Nick Anstead and Ben O'Loughlin. 2011. The emerging viewertariat and BBC question time: Television debate and real-time commenting online. *The international journal of press/politics* 16, 4 (2011), 440–462. https://doi.org/10.1177/1940161211415519

[5] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. https://doi.org/10.48550/arXiv.2303.00747

[6] Tetiana Bas. 2024. Assessing Gender Bias in LLMs: Comparing LLM Outputs with Human Perceptions and Official Statistics.

[7] William L. Benoit. 2016. Political Election Debates. In *The International Encyclopedia of Political Communication*. Wiley Blackwell, Malden, MA, USA and Oxford, UK, 1–10. https://doi.org/10.1002/9781118541555.wbiepc127

[8] William L. Benoit and Glenn J. Hansen. 2004. Presidential Debate Watching, Issue Knowledge, Character Evaluation, and Vote Choice. *Human Communication Research* 30, 1 (2004), 121–144. https://doi.org/10.1111/j.1468-2958.2004.tb00727.x

[9] William L. Benoit, Glenn J. Hansen, and Rebecca M. Verser. 2003. A meta-analysis of the effects of viewing U.S. presidential debates. *Communication Monographs* 70, 4 (2003), 335–350. https://doi.org/10.1080/0363775032000179133

[10] William L. Benoit, Mitchell S. McKinney, and R. Lance Holbert. 2001. Beyond learning and persona: extending the scope of presidential debate effects. *Communication Monographs* 68, 3 (2001), 259–273. https://doi.org/10.1080/03637750128060

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[12] Guilherme Carneiro, Miguel Nacenta, Alice Toniolo, Gonzalo Mendez, and Aaron J Quigley. 2019. Deb8: A tool for collaborative analysis of video. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*. Association for Computing Machinery, New York, NY, USA, 47–58. https://doi.org/10.1145/3317697.3323358

[13] Pew Research Center. 2024. Changing Partisan Coalitions in a Politically Divided Nation, Chapter 4. Age, generational cohorts and party identification. https://www.pewresearch.org/politics/2024/04/09/age-generational-cohorts-and-party-identification/ Accessed: 2024-10-06.

[14] Andrew Chadwick, Ben O'Loughlin, and Cristian Vaccari. 2017. Why People Dual Screen Political Debates and Why It Matters for Democratic Engagement. *Journal of Broadcasting & Electronic Media* 61, 2 (2017), 220–239. https://doi.org/10.1080/08838151.2017.1309415

[15] Teresa Chambel, Maiur N. Chhaganlal, and Luís A. R. Neng. 2011. Towards immersive interactive video through 360° hypervideo. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE '11)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/2071423.2071518

[16] Yen-Shan Chen, Jing Jin, Peng-Ting Kuo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Llms are biased evaluators but not biased for retrieval augmented generation.

[17] Zhutian Chen, Qisen Yang, Jiarui Shan, Tica Lin, Johanna Beyer, Haijun Xia, and Hanspeter Pfister. 2023. iBall: Augmenting Basketball Videos with Gaze-moderated Embedded Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581266

[18] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios.

[19] Stephen Coleman. 2020. Introduction: Televised Election Debates as Spectacle and Reflection. In *Routledge international handbook on electoral debates*. Routledge, London, UK and New York, NY, USA, 5–14.

[20] NetMQ contributors. 2024. NetMQ: A native C# implementation of ZeroMQ for .NET. https://github.com/zeromq/netmq.

[21] Jasmine Cox, Rhianne Jones, Chris Northwood, Jonathan Tutcher, and Ben Robinson. 2017. Object-Based Production: A Personalised Interactive Cooking Application. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 79–80. https://doi.org/10.1145/3084289.3089912

[22] Anna De Liddo, Brian Plüss, and Paul Wilson. 2017. A Novel Method to Gauge Audience Engagement with Televised Election Debates through Instant, Nuanced Feedback Elicitation. In *Proceedings of the 8th International Conference on Communities and Technologies (C&T '17)*. Association for Computing Machinery, New York, NY, USA, 68–77. https://doi.org/10.1145/3083671.3083673

[23] Anna De Liddo, Nieves Pedreira Souto, and Brian Plüss. 2021. Let's replay the political debate: Hypervideo technology for visual sensemaking of televised election debates. *International Journal of Human-Computer Studies* 145 (2021), 102537. https://doi.org/10.1016/j.ijhcs.2020.102537

[24] James N. Druckman. 2003. The Power of Television Images: The First Kennedy-Nixon Debate Revisited. *The Journal of Politics* 65, 2 (2003), 559–571. https://doi.org/10.1111/1468-2508.t01-1-00015

[25] Evelien D'heer and Cédric Courtois. 2016. The changing dynamics of television consumption in the multimedia living room. *Convergence* 22, 1 (2016), 3–17. https://doi.org/10.1177/1354856514543451

[26] Tom Feltwell, Gavin Wood, Kiel Long, Phillip Brooker, Tom Schofield, Ioannis Petridis, Julie Barnett, John Vines, and Shaun Lawson. 2017. "I've been manipulated!": Designing Second Screen Experiences for Critical Viewing of Reality TV. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2252–2263. https://doi.org/10.1145/3025453.3025833

[27] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. YOLOX: Exceeding YOLO Series in 2021.

[28] David Geerts, Rinze Leenheer, Dirk De Grooff, Joost Negenman, and Susanne Heijstraten. 2014. In front of and behind the second screen: viewer and producer perspectives on a companion app. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '14)*. Association for Computing Machinery, New York, NY, USA, 95–102. https://doi.org/10.1145/2602299.2602312

[29] Katerina Gorkovenko and Nick Taylor. 2016. Politics at Home: Second Screen Behaviours and Motivations During TV Debates. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2971485.2971514

[30] Katerina Gorkovenko and Nick Taylor. 2017. Understanding how people use twitter during election debates. In *HCI '17: Proceedings of the 31st British Computer Society Human Computer Interaction Conference* (Sunderland, UK) *(HCI '17)*. BCS Learning & Development Ltd., Swindon, GBR, Article 88, 6 pages. https://doi.org/10.14236/ewic/HCI2017.88

[31] Katerina Gorkovenko and Nick Taylor. 2019. Audience and Expert Perspectives on Second Screen Engagement with Political Debates. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '19)*. Association for Computing Machinery, New York, NY, USA, 70–82. https://doi.org/10.1145/3317697.3323352

[32] Katerina Gorkovenko and Nick Taylor. 2019. Audience and Expert Perspectives on Second Screen Engagement with Political Debates. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '19)*. Association for Computing Machinery, New York, NY, USA, 70–82. https://doi.org/10.1145/3317697.3323352

[33] Katerina Gorkovenko, Nick Taylor, and Jon Rogers. 2017. Social Printers: A Physical Social Network for Political Debates. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2269–2281. https://doi.org/10.1145/3025453.3025611

[34] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation.

[35] Michael E. Holmes, Sheree Josephson, and Ryan E. Carney. 2012. Visual attention to television programs with a second-screen application. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 397–400. https://doi.org/10.1145/2168556.2168646

[36] Jonathan Hook. 2018. Facts, Interactivity and Videotape: Exploring the Design Space of Data in Interactive Video Storytelling. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '18)*. Association for Computing Machinery, New York, NY, USA, 43–55. https://doi.org/10.1145/3210825.3210826

[37] Michael Hout and Christopher Maggio. 2021. Immigration, Race & Political Polarization. *Daedalus* 150, 2 (2021), 40–55. https://doi.org/10.1162/daed_a_01845

[38] Parag Jain, Andreea Marzoca, and Francesco Piccinno. 2024. Structsum Generation for Faster Text Comprehension. https://doi.org/10.48550/arXiv.2401.06837

[39] Val Jones, Weston Gaylord, Ari Palitz, Inge Auerbacher, Andrea Csabai, Dan Franke, Moira Hamilton, Kia Hays, Andrew Lush, Laiylaly Mandujano, Ruxandra Popescu, Francis Rockwell, Zoe Roellin, Ronen Shaham, and Stephen Smith. 2023. Tell Me, Inge…: An Interactive Interview with a Holocaust Survivor. In *ACM SIGGRAPH 2023 Real-Time Live! (SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3588430.3597256

[40] Kyle Jorgensen, Zhiqun Zhao, Haohong Wang, Mea Wang, and Zhihai He. 2021. Context-Aware Question-Answer for Interactive Media Experiences. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences (IMX '21)*. Association for Computing Machinery, New York, NY, USA, 156–166. https://doi.org/10.1145/3452918.3458795

[41] I Kazanidis, G Palaigeorgiou, A Papadopoulou, and A Tsinakos. 2018. Augmented Interactive Video: Enhancing Video Interactivity for the School Classroom.

[42] Daniel V. Klein and Paul McCartney. 2024. Delivering Live Q&A in Videos via Synthetic Content Generation Using Generative Artificial Intelligence. https://www.tdcommons.org/dpubs_series/6984 Accessed: 2024-05-08.

[43] Jon A Krosnick and Stanley Presser. 2010. Question and Questionnaire Design. In *Handbook of Survey Research* (2 ed.), Peter V. Marsden and James D. Wright (Eds.). Emerald, Bingley, UK, [268–274].

[44] T. S. Desai and D. C. Kulkarni. 2022. Assessment of Interactive Video to Enhance Learning Experience: A Case Study. https://doi.org/10.16920/jeet/2022/v35is1/22011

[45] Tica Lin, Chen Zhu-Tian, Yalong Yang, Daniele Chiappalupi, Johanna Beyer, and Hanspeter Pfister. 2023. The Quest for Omnioculars: Embedded Visualization for Augmenting Basketball Game Viewing Experiences. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 962–972. https://doi.org/10.1109/TVCG.2022.3209353

[46] Carles Marín Lladó and José Manuel Pérez Tornero. 2020. Un modelo conceptual para analizar los debates electorales en TV. Mediatización y ceremonias televisivas. , 229–245 pages. https://doi.org/10.4185/RLCS-2020-1445

[47] Celina Ma, Haohong Wang, Hao Sun, Elliot van Huijgevoort, Mea Wang, and Zhihai He. 2021. Powering TV Experiences with Anytime Environmental Exploration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3411763.3451600

[48] Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

[49] Mitchell S. McKinney and Sumana Chattopadhyay. 2007. Political Engagement Through Debates: Young Citizens' Reactions to the 2004 Presidential Debates. *American Behavioral Scientist* 50, 9 (2007), 1169–1182. https://doi.org/10.1177/0002764207300050 Publisher: SAGE Publications Inc.

[50] Mitchell S. McKinney, J. Brian Houston, and Joshua Hawthorne. 2014. Social Watching a 2012 Republican Presidential Primary Debate. *American Behavioral Scientist* 58, 4 (2014), 556–573. https://doi.org/10.1177/0002764213506211

[51] Mitchell S. McKinney and Leslie A. Rill. 2009. Not Your Parents' Presidential Debates: Examining the Effects of the CNN/YouTube Debates on Young Citizens' Civic Engagement. *Communication Studies* 60, 4 (2009), 392–406. https://doi.org/10.1080/10510970903110001

[52] Leila Methnani, Andrea Aler Tubella, Virginia Dignum, and Andreas Theodorou. 2021. Let Me Take Over: Variable Autonomy for Meaningful Human Control. *Frontiers in Artificial Intelligence* 4 (2021), 737072. https://doi.org/10.3389/frai.2021.737072

[53] Timothy Neate, Matt Jones, and Michael Evans. 2016. Interdevice Media: Choreographing Content to Maximize Viewer Engagement. *Computer* 49, 12 (Dec. 2016), 42–49. https://doi.org/10.1109/MC.2016.375 Conference Name: Computer.

[54] Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas J. Wood. 2020. Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. *Political Behavior* 42, 3 (2020), 939–960. https://doi.org/10.1007/s11109-019-09528-x

[55] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities.

[56] Brian Plüss and Anna De Liddo. 2018. Democratic replay: Enhancing tv election debates with interactive visualisations.

[57] Brian Plüss and Anna De Liddo. 2015. Engaging Citizens with Televised Election Debates through Online Interactive Replays. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '15)*. Association for Computing Machinery, New York, NY, USA, 179–184. https://doi.org/10.1145/2745197.2755521

[58] Carolyn C Preston and Andrew M Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104, 1 (2000), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

[59] R. Priyakanth, Radha Abburi, and Manne Praveena. 2020. Design and Impact of Interactive Video Content for the Improvement of Student Engagement and Learning. *Journal of Engineering Education Transformations* 34, Special Issue (2020), 518–523. https://doi.org/10.16920/jeet/2021/v34i0/157204

[60] Weina Ran and Masahiro Yamamoto. 2019. Media multitasking, second screening, and political knowledge: Task-relevant and task-irrelevant second screening during election news consumption. *Journal of Broadcasting & Electronic Media* 63, 1 (2019), 1–19. https://doi.org/10.1080/08838151.2019.1565659

[61] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084

[62] Anna Marie Rezk and Mads Haahr. 2020. The Case for Invisibility: Understanding and Improving Agency in Black Mirror's Bandersnatch and Other Interactive Digital Narrative Works. In *Interactive Storytelling*, Anne-Gwenn Bosser, David E. Millard, and Charlie Hargood (Eds.). Springer International Publishing, Cham, 178–189. https://doi.org/10.1007/978-3-030-62516-0_16

[63] Anna Marie Rezk and Mads Haahr. 2022. Beyond free will: Understanding approaches to agency and their suitability for Bandersnatch-like titles. *Entertainment Computing* 43 (2022), 100500. https://doi.org/10.1016/j.entcom.2022.100500

[64] Christian Roth and Hartmut Koenitz. 2019. Bandersnatch, Yea or Nay? Reception and User Experience of an Interactive Digital Narrative Video. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '19)*. Association for Computing Machinery, New York, NY, USA, 247–254. https://doi.org/10.1145/3317697.3325124

[65] Robert C. Rowland. 2021. The 2020 presidential debates: reasoned argument or political theater. *Argumentation and Advocacy* 57, 3-4 (2021), 218–235. https://doi.org/10.1080/10511431.2021.1949535

[66] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. https://doi.org/10.48550/arXiv.2304.11406 arXiv:2304.11406 [cs].

[67] Gabi Schaap, Mariska Kleemans, and Anna Van Cauwenberge. 2018. Second screening for news: Effects of presentation on information processing and program liking. *Computers in Human Behavior* 84 (July 2018), 76–85. https://doi.org/10.1016/j.chb.2018.02.006

[68] Vinay Setty. 2024. FactCheck Editor: Multilingual Text Editor with End-to-End fact-checking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2744–2748. https://doi.org/10.1145/3626772.3657663

[69] Dhavan V. Shah, Alex Hanna, Erik P. Bucy, David S. Lassen, Jack Van Thomme, Kristen Bialik, JungHwan Yang, and Jon C. W. Pevehouse. 2016. Dual Screening During Presidential Debates: Political Nonverbals and the Volume and Valence of Online Expression. *American Behavioral Scientist* 60, 14 (2016), 1816–1843. https://doi.org/10.1177/0002764216676245

[70] Catharyn C. Shelton, Annie E. Warren, and Leanna M. Archambault. 2016. Exploring the Use of Interactive Digital Storytelling Video: Promoting Student Engagement and Learning in a University Hybrid Course. *TechTrends* 60, 5 (2016), 465–474. https://doi.org/10.1007/s11528-016-0082-z

[71] Pedro Silva, Yasmin Amer, William Tsikerdanos, Jesse Shedd, Isabel Restrepo, and Janet Murray. 2015. A Game of Thrones Companion: Orienting Viewers to Complex Storyworlds via Synchronized Visualizations. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '15)*. Association for Computing Machinery, New York, NY, USA, 167–172. https://doi.org/10.1145/2745197.2755519

[72] Simen Hunding Strømme, Chris Ronald Hermansen, and Lene Hegland. 2025. Derfor har vi laget KI-Kjetil. https://info.tv2.no/info/artikkel/derfor-har-vi-laget-ki-kjetil Accessed: 2025-01-17.

[73] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. https://doi.org/10.1145/3613904.3642513

[74] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, Piscataway, NJ, USA, 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962 ISSN: 2381-8549.

[75] Ying Xu, Valery Vigil, Andres S. Bustamante, and Mark Warschauer. 2022. "Elinor's Talking to Me!":Integrating Conversational AI into Children's Narrative Science Programming. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3491102.3502050

[76] Chance York, James D. Ponder, Zach Humphries, Catherine Goodall, Michael Beam, and Carrie Winters. 2020. Effects of Fact-Checking Political Misinformation on Perceptual Accuracy and Epistemic Political Efficacy. *Journalism &*

Mass Communication Quarterly 97, 4 (2020), 958–980. https://doi.org/10.1177/1077699019890119
[77] Hang Yuan, Zhongyue Che, Shao Li, Yue Zhang, Xiaomeng Hu, and Siyang Luo. 2024. The high dimensional psychological profile and cultural bias of ChatGPT.
[78] Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models.
[79] Jian-Hua Zhu, J Ronald Milavsky, and Rahul Biswas. 1994. Do televised debates affect image perception more than issue knowledge? A study of the first 1992 presidential debate. *Human Communication Research* 20, 3 (1994), 302–333.
[80] Fred R. H. Zijlstra. 1993. *Efficiency in Work Behavior: A Design Approach for Modern Tools.* Ph. D. Dissertation. Delft University of Technology, Delft, Netherlands.

A Interactive Features Questionnaire

Date:

Participant:

Function: Fact Checker

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This function is relevant to my needs as a debate viewer.	○	○	○	○	○	○	○
This function led to a more engaging experience.	○	○	○	○	○	○	○
The feedback I received clarified the background information, circumstances, and/or setting that surround the topic being discussed.	○	○	○	○	○	○	○
This function effectively supplemented the video content, enhancing my overall understanding of the debate topics.	○	○	○	○	○	○	○
I would use this function when viewing political debates in the future.	○	○	○	○	○	○	○
Overall, I was satisfied with this function.	○	○	○	○	○	○	○

Function: Platform Checker

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This function is relevant to my needs as a debate viewer.	○	○	○	○	○	○	○
This function led to a more engaging experience.	○	○	○	○	○	○	○
The feedback I received clarified the background information, circumstances, and/or setting that surround the topic being discussed.	○	○	○	○	○	○	○
This function effectively supplemented the video content, enhancing my overall understanding of the debate topics.	○	○	○	○	○	○	○
I would use this function when viewing political debates in the future.	○	○	○	○	○	○	○
Overall, I was satisfied with this function.	○	○	○	○	○	○	○

Function: More Information

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This function is relevant to my needs as a debate viewer.	○	○	○	○	○	○	○
This function led to a more engaging experience.	○	○	○	○	○	○	○
The feedback I received clarified the background information, circumstances, and/or setting that surround the topic being discussed.	○	○	○	○	○	○	○
This function effectively supplemented the video content, enhancing my overall understanding of the debate topics.	○	○	○	○	○	○	○
I would use this function when viewing political debates in the future.	○	○	○	○	○	○	○
Overall, I was satisfied with this function.	○	○	○	○	○	○	○

Function: Stance Checker

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This function is relevant to my needs as a debate viewer.	○	○	○	○	○	○	○
This function led to a more engaging experience.	○	○	○	○	○	○	○
The feedback I received clarified the background information, circumstances, and/or setting that surround the topic being discussed.	○	○	○	○	○	○	○
This function effectively supplemented the video content, enhancing my overall understanding of the debate topics.	○	○	○	○	○	○	○
I would use this function when viewing political debates in the future.	○	○	○	○	○	○	○
Overall, I was satisfied with this function.	○	○	○	○	○	○	○

Function: Explore Topic

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This function is relevant to my needs as a debate viewer.	○	○	○	○	○	○	○
This function led to a more engaging experience.	○	○	○	○	○	○	○
The feedback I received clarified the background information, circumstances, and/or setting that surround the topic being discussed.	○	○	○	○	○	○	○
This function effectively supplemented the video content, enhancing my overall understanding of the debate topics.	○	○	○	○	○	○	○
I would use this function when viewing political debates in the future.	○	○	○	○	○	○	○
Overall, I was satisfied with this function.	○	○	○	○	○	○	○

Function: Opinion Poll

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This function is relevant to my needs as a debate viewer.	○	○	○	○	○	○	○
This function led to a more engaging experience.	○	○	○	○	○	○	○
This function helped in expressing my feelings towards the debate.	○	○	○	○	○	○	○
This function encouraged me to reflect on my viewpoints.	○	○	○	○	○	○	○
The visuals complemented the information delivered by this function.	○	○	○	○	○	○	○
I would use this function when viewing political debates in the future.	○	○	○	○	○	○	○
Overall, I was satisfied with this function.	○	○	○	○	○	○	○

Function: React

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This function is relevant to my needs as a debate viewer.	○	○	○	○	○	○	○
This function led to a more engaging experience.	○	○	○	○	○	○	○
This function helped in expressing my feelings towards the debate.	○	○	○	○	○	○	○
This function encouraged me to reflect on my viewpoints.	○	○	○	○	○	○	○
The visuals complemented the information delivered by this function.	○	○	○	○	○	○	○
I would use this function when viewing political debates in the future.	○	○	○	○	○	○	○
Overall, I was satisfied with this function.	○	○	○	○	○	○	○

B Post-Viewing Questionnaire

Date:

Participant:

Mode:

	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This viewing method assisted me to critically reflect on the candidates discussion.	◯	◯	◯	◯	◯	◯	◯
This viewing method assisted in my understanding of each candidates viewpoint.	◯	◯	◯	◯	◯	◯	◯
This viewing method assisted in linking the candidate perspectives with those of their respective political parties.	◯	◯	◯	◯	◯	◯	◯
This viewing method helped me evaluate the credibility of the candidates' statements.	◯	◯	◯	◯	◯	◯	◯
I was able to easily find the information I needed with this viewing method.	◯	◯	◯	◯	◯	◯	◯
Navigating the information sources was intuitive with this viewing method.	◯	◯	◯	◯	◯	◯	◯
Finding and viewing additional information did not overly distract from watching the debate.	◯	◯	◯	◯	◯	◯	◯
The viewing experience was engaging using this viewing method.	◯	◯	◯	◯	◯	◯	◯
I found the debate entertaining to watch with this viewing method.	◯	◯	◯	◯	◯	◯	◯
I felt immersed in the political debate with this viewing method.	◯	◯	◯	◯	◯	◯	◯
This viewing method provided sufficient explanations of the debate topics.	◯	◯	◯	◯	◯	◯	◯
This viewing method assisted in clarifying any uncertainties I had during the debate.	◯	◯	◯	◯	◯	◯	◯
While viewing I was able to easily identify the positions of each candidate on key issues.	◯	◯	◯	◯	◯	◯	◯
This viewing method provided enough context to enhance my understanding.	◯	◯	◯	◯	◯	◯	◯
This viewing method effectively connected the debate topics to broader social, political, or historical contexts.	◯	◯	◯	◯	◯	◯	◯
This viewing method effectively clarified complex terms and concepts used within the debate.	◯	◯	◯	◯	◯	◯	◯
Given the choice, I would watch political debates with this viewing method in the future.	◯	◯	◯	◯	◯	◯	◯
I would recommend this debate viewing method to friends and family.	◯	◯	◯	◯	◯	◯	◯
I was satisfied with this debate viewing experience.	◯	◯	◯	◯	◯	◯	◯

1

[4] This questionnaire uses the term "method" for participant clarity; however, in the main text, we refer to it as "mode."