Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Fakultät für Mathematik und Informatik
Institut für Informatik
Mensch-Computer-Interaktion

Bachelorarbeit

zur Erlangung des Grades
Bachelor of Science (B.Sc.)
im Studiengang Mensch-Computer-Systeme
an der Universität Würzburg

Leveraging Summarizations from LLMs for enhancing Community Online Deliberation Tools

vorgelegt von
Jan Steinhauer
Matrikelnummer: 2733107

am June 8, 2024

Betreuer/Prüfer:
Dr. Sebastian Oberdörfer, Informatik IX, Universität Würzburg

Abstract

This study explores the application of large language models to enhance online community deliberation by investigating various summarization techniques. The research specifically evaluates six distinct methods: summaries with citations, simple language summaries, participant-initiated summaries, tree-structured summaries, graph summaries, and the summary with the ability to chat. Utilizing the digital community deliberation platform Polis, the study uses a within-subject design experiment to evaluate user interactions and preferences across these techniques.

Participants generally preferred summaries that included citations for added credibility, while tree-structured summaries were favored for their interactive depth and organization. Simple language summaries were appreciated for their clarity but criticized for lacking comprehensive detail. Participant-initiated summaries. The summary with the ability to chat was found to be user-friendly and intuitive, though concerns about the reliability of LLM-generated responses were addressed by participants. Graph summaries provided valuable insights.

The findings suggest that combining features, such as integrating tree-structured summaries with citations or improving graph summaries with interactive elements, could offer a more robust and user-friendly experience. The study concludes that LLM summaries have significant potential to facilitate more meaningful and informed online community deliberations, with recommendations for future research to explore optimal design and integration strategies.

Contents

List of Figures

1. Introduction

"Public Deliberation, simply defined, is the discussion and choice-making that is necessary before we can solve problems that affect our communities together." This is the definition of public deliberation by the University of Houston (University of Houston-Downtown Center for Public Deliberation, 2024). Throughout history, community deliberation has always been something that people have engaged with. It is as old as the civilizations itself. Classical Athens engaged in public deliberations to make a decision on whether to exile people or not (Robertson, 2023). Also, ancient Rome utilized community deliberation to create laws that apply to every citizen, no matter the class or wealth (Britannica, 2020). This and numerous more essential decisions were made with community deliberation.

Through the study "Future directions for public deliberation" (Levine et al., 2005), it was found that there is an incontiguous interest in community deliberation throughout history. Depending on the circumstances, people are drawn more and sometimes less, to a public discussion. Historically speaking, community deliberation has seen great interest in U.S. history during the Progressive Era and the Great Depression (Levine et al., 2005). The people who engage in community deliberations are often a wide range of participants with various backgrounds. The participants of public deliberation benefit greatly from engaging in these discussions. First, it leads to high-quality decisions that benefit the community. Furthermore, people feel deeply satisfied when participating in community deliberation. This can often lead to repeated involvement (Levine et al., 2005). Even though community deliberation has numerous advantages, it also faces challenges. For example, achieving full consensus in a discussion is rare. Therefore, voting is often needed to come to a decision. Moreover, participants can struggle to misunderstand the current topic, which can lead to a wrong conclusion.

Nowadays, community deliberation often takes place digitally. Community deliberation on the Internet can change how people interact and discuss with each other. A digital community deliberation platform has many of the same advantages and

disadvantages as public in-person deliberation, for example, the cases stated above. Furthermore, hosting a community deliberation online can oppose further obstacles and benefits. One disadvantage is that the quality of the discussion could be significantly impacted by the design of the digital deliberation platform (Shortall et al., 2022). On the other hand, having a public online deliberation can help with accessibility and inclusivity. Online deliberation can engage a wider audience by eliminating geographical constraints, enabling individuals from various locations to participate without the necessity of travel. This approach is particularly beneficial for those who lack the resources or capability to attend in-person meetings (Bobbio, 2019).

There have been numerous attempts to improve online deliberation platforms. One of them was to implement a design that was more structural and offered visualizations. It has shown that features that facilitate better argument visualization and navigation of discussions conclude in allowing participants to engage meaningfully without getting lost in disorganized threads (Yeo et al., 2024). Furthermore, Studies have shown that prompts encouraging users to reflect on their own and others' viewpoints can reduce bias and promote more thoughtful and inclusive discussions (Yeo et al., 2024). Moreover, attempts have been made to incorporate large language model (LLM) summaries into online public deliberation platforms in order to give the participants a better experience. The participants should gain a deeper understanding of the discussion topic and of the discussion itself through a summary. Through the platform ConsiderIt, researchers have found out that it helps users of online deliberation platforms to have a summary of the pro and con arguments of the discussion (Kriplean et al., 2012). Other papers have found that public deliberation platforms benefit from a tree-structured summary. In this case, participants first get a broad summary, which is generated by an LLM. Within this summary, users can generate further summaries to understand each subtopic. This design allows the users to branch deeper and deeper into the different subfields (Zhang et al., 2017). Moreover, researchers have evaluated that LLM can help deliberation platforms like polis with topic modeling, summarization, finding group identity and consensus, and vote prediction (Small et al., 2023). When researchers created a prompt to summarize the discussion on the polis platforms, they only explored one technique. Besides that, there are numerous different summarization techniques that can help the user understand the topic and the discussion better, which leads to better and more well-informed decisions.

Therefore, this paper is going to investigate what are good summarization techniques for LLMs to sum up discussions on digital community deliberation tools. First, the theoretical background knowledge will be discussed. This consists of an explanation of the digital community deliberation platform polis as well as unmanned area mobility. After that, the paper will elaborate on the refinement of the prompt technique. Afterward, the Hypothesis will be described. This is followed by an explanation of the implementation. Hereby, the implementation of the UI changes will be described, as well as the prompt implementation. In the next step, the mythology will be illustrated. Subsequently, the result will be discussed. The following section is going to be the discussion. The thesis will be concluded by talking about possibilities for future work.

2. Theoretical background knowledge

2.1. Polis Platform

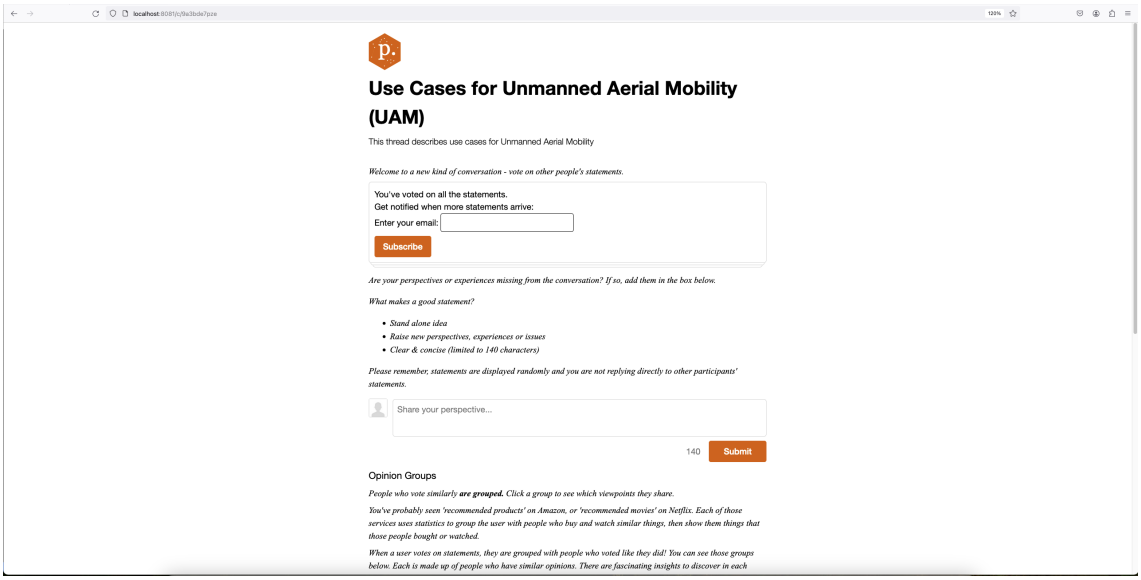

Figure 2.1.. Polis Page

The digital community deliberation platform Polis ("Polis," 2024) is a tool designed to facilitate large-scale, structured discussions and decision-making processes online. It enables participants to engage in meaningful deliberation and reach a consensus on complex issues. Polis operates by collecting and analyzing the opinions of participants through a simple interface. Users can submit their own statements and vote on the statements of others by either agreeing, disagreeing, or skipping them. Polis uses algorithms to identify patterns in the responses, clustering participants into opinion groups based on their voting behavior. This process helps visualize the landscape of opinions, highlighting areas of agreement and contention.

Polis has been effectively utilized in various real-world scenarios to facilitate public deliberation and decision-making. For instance, the government of Taiwan employed Polis to gather public opinion on several contentious issues, such as the regulation

of ride-sharing services and the implementation of same-sex marriage legislation. Through the platform, thousands of citizens were able to participate in the discussion, submitting their views and voting on others' statements (Staff, 2024).

Overall, Polis represents a significant advancement in the field of digital community deliberation. It enables a wide range of participants to engage in structured and informed discussions. Polis helps communities make better decisions and solve complex problems together.

2.2. Polis Graph

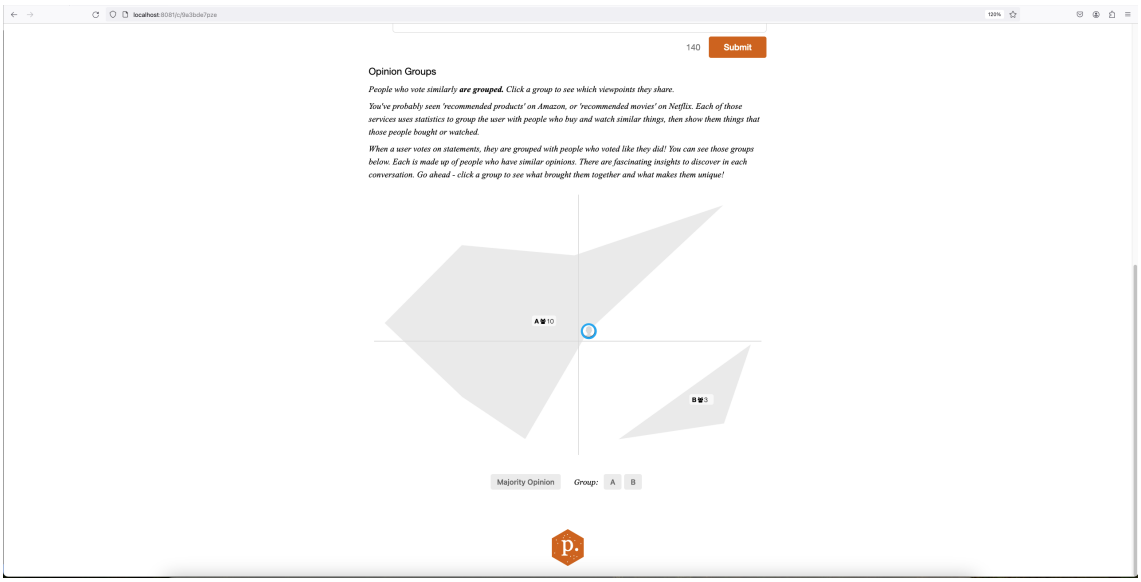

Figure 2.2.. Polis Graph Example

The Polis platform graph is an interactive visualization tool that groups people based on their similar voting patterns on various statements, depicted as distinct polygons on a central axes graph. Each group's position and shape on the graph indicate their collective opinions, with users able to click on these groups to explore shared viewpoints. The platform's interactivity allows users to view majority opinions on specific statements by selecting the group and statement number. Summaries of each group's major trends and disagreements are displayed alongside the graph, offering detailed insights into their stances. By comparing agreement or disagreement percentages on statements across different groups, the graph facilitates a deeper understanding of public opinion, highlighting areas of consensus or contention.

2.3. Unmanned Area Mobility

During the study, participants engaged in discussions about unmanned area mobility (UAM). The following paragraph should give a short overview of UAMs. Unmanned Air Mobility refers to the use of drones and other unmanned aerial systems (UAS) to transport goods and people. This technology promises significant benefits such as reducing urban traffic congestion, offering faster delivery services, and improving accessibility to remote areas. It involves various applications, including delivery of goods, emergency response, and urban air taxis. The development and integration of these systems require robust regulatory frameworks to ensure safety, security, and environmental sustainability (Agency, 2024).

2.4. One-Shot Learning

One-shot learning in large language models signifies that the models can understand new information or tasks from just a single example. Unlike traditional models that need lots of data, LLMs use their training on a wide range of texts to learn from very little input. Thanks to this ability, LLMs can transfer learning and pattern recognition to new tasks (Lark, 2024).

2.5. Prompt Tweaking

Prompt tweaking means modifying the input given to a machine learning model, for example, a large language model, to achieve the desired response or behavior. This process involves altering the phrasing, incorporating additional context, or re-organizing the prompt. These adjustments make the input clearer and more precise, resulting in enhanced accuracy and relevance in the model's output (Brown et al., 2020).

3. Prompt Techniques

In this section of the paper, I will first talk about why GPT 3.5 Turbo was chosen for the LLM of the experiment. Afterward, each of the six independent variables will explain why they were chosen and how they were implemented.

3.1. Selection of Model GPT 3.5 Turbo

Currently, there are numerous large language models on the market that can summarize texts. For example, Google released Gemini, Meta developed Lama, Openai created GPT, etc. According to paper Watanangura et al. (2023), the model that is best suited for the summary would be GPT. It was described that Chat GPT combines the benefits of both extractive and abstractive techniques, allowing it to effectively grasp the major viewpoints of the text while remaining natural and coherent. Due to the described properties, GPT was chosen for the experiment. Within the Openai family, the GPT Model 3.5 Turbo was selected because it was the most advanced model at the time and had a public access API.

3.2. Prompt Summary with Citation

Providing citations within a text has been shown in multiple scientific papers to improve credibility. For example, Watanangura et al. (2023) found out that comments in health-focused online discussion groups were viewed with a higher credibility if citations were provided. Furthermore, it allows the user to gain a deeper understanding of the topic by reading further into the issue. The following prompt was used to generate a summary of the online discussion A.1. The first sentence of the prompt is responsible for creating the summary. The second sentence handles that the LLM provides scientific citation. Through multiple iterations, it was found that the phrase "Important: " needed to be included. Sometimes the LLM would

not provide scientific links if "Important: " wasn't included. The last sentence of the prompt is responsible that the link is clickable. By giving the LLM the order to highlight the links in the markup annotation, GPT 3-5 turbo returns a clickable hyperlink link. The described prompt technique consistently delivers a summary of the discussion topic, with links at the end to scientific papers, which are clickable.

3.3. Prompt Summary for Simple Language

Offering a summary, which is written in simple language, has numerous benefits. It often results in a better understanding of the user. Furthermore, text with simple language is more accessible to people whose English is not their mother tongue. For the study of this paper, the following prompt was used A.2. It was decided that the prompt should ask for a summary suitable for a 16-year-old. The age was determined by trying different prompts and evaluating them. We also took into consideration scientific evidence of the ability to comprehend at different age groups.

First, we tried ages 14, 16, and 18 years. If the prompt age was 14, the resulting response was typically short. The LLM tried to explain the concepts only with very high-level metaphors. When the summary was suitable for a 16-year-old, the result included high-level metaphors and a few technical terms. Length-wise, the summary should be around four to five sentences. A summary for an 18-year-old would lead to a more extended summary. The text would include mainly technical terms and explain more complex details. Upon evaluating the trial and error process, we started fancying the prompt, which is suitable for a 16-year-old.

Scientifically speaking, the age for the prompt also made sense because 16-year-olds are usually in their mid-adolescence, where they have the ability to handle abstract topics. Also, they start to understand complex relations (American Academy of Child Adolescent Psychiatry, 2016). For both of the reasons stated above, a summary was created that is suitable for a 16-year-old.

3.4. Prompt Summary for Participant Initiated

For the participant-initiated summary the following prompt was used A.3. The large language model was given the information on the title, subtitle, and all of the com-

ments. Furthermore, the prompt was specified to summarize the main arguments of the discussion, to identify common themes as well as return a comprehensive overview that can be quickly understood by the user. Through multiple iterations, it was found that this prompt offers a proper understanding of the topic and the discussion.

3.5. Prompt Summary for the Tree Structure

The design of the tree summary was primarily inspired by the Wikium research paper (Zhang et al., 2017). In the tree structure, users are initially presented with a summary of the discussion. Below it, three buttons display the main topics of the discussion. When a user clicks on a button, a new summary of the selected topic is generated and appears at the end of the previous summary.

The tree summary method consists of four prompts. The first prompt summarizes the discussion, the second prompt identifies the three main topics of the discussion, the third prompt generates further summaries of the main topics, and the fourth prompt extracts three main topics from these further summaries.

For the first prompt, the LLM received the topic and description of the discussion along with all the statements. It was then prompted to summarize the discussion, determine the primary arguments or viewpoints, and provide a comprehensive overview. This prompt was refined through trial and error, and its final version can be found in the A.4.1.

To identify the three main topics of the discussion, the LLM was given the topic, description, and statements of the discussion. It was then prompted to return the topics in the following schema: [topic1] [topic2] [topic3]. This prompt structure, detailed in the A.4.2, was found to be the most effective through multiple iterations. By explicitly specifying the schema, the LLM almost always returned the topics in the required format, which was essential for displaying the results as buttons. Ensuring the LLM consistently followed the [topic1] [topic2] [topic3] schema required one-shot learning.

The prompt for generating further summaries uses the same structure as the first prompt. However, instead of the discussion statements, it uses the name of the main topic clicked by the user. This prompt is also available in the A.4.3.

The final prompt for the tree summary follows a similar logic to the second prompt, but instead of discussion statements, it receives the name of the main topic (A.4.4).

The prompts that generated the three main topics utilized one-shot learning. The message sent to OpenAI consisted of two parts: the one-shot learning portion and the prompt itself. As explained in the theoretical background, one-shot learning involves showing the LLM a prompt and the corresponding expected response. Through trial and error, it was determined that one-shot learning was crucial for ensuring the LLM consistently returned the three main topics in the intended schema.

3.6. Prompt Graph Summary

Upon first sight of the polis graph, it wasn't obvious how to read it. This was a problem that not only I have encountered but also people on the internet (Jackson, 2022). The reasons for that are that the axis is not labeled, it has a polygon shape, etc. To help users understand the graph, we came up with the perfect prompt. In order to find the perfect prompt to summarize the Polis graph, multiple approaches were made. Hereby, we compared prompts, which utilized data from the graph object, as well as the raw data, and also the approach from the paper Small et al., 2023. After evaluating all prompt it was tried to improve the result by tweaking the prompt. Through multiple interations the prompt that was best for summarizing the graph was found.

To find this ideal prompt for summarizing the polis graph, various approaches were tested. We compared prompts that used data from the graph object, the raw data, and a method from the paper by Small Small et al., 2023. After evaluating these prompts, we refined them through multiple iterations. Ultimately, the best prompt for summarizing the graph was identified.

3.6.1. Evaluating LLM Graph Summaries with different Datapoints

In the initial brainstorming sessions on creating LLM generated summaries for the Polis graph, we focused on identifying the most relevant data points that the LLM could use. Three primary sources of data points were identified. First, the Polis platform provided a JSON graph object, which stored numerous variables characterizing

the graph. For example, it contained statements, comment representativeness, and group consensus. Secondly, we considered providing the LLM with the raw statements and voting data for summary generation. Lastly, we attempted to replicate the prompts from the paper "Opportunities and Risks of LLMs for Scalable Deliberation with Polis" (Small et al., 2023). We then compared all the described summary approaches to determine the most effective method.

In the first iteration, after analyzing the JSON graph object, we created eight prompts containing various combinations of the following information: base cluster, comment priority, group-aware consensus, group clusters, principal components analysis, representativeness of comments, and combinations of these variables. All prompts followed a similar structure. It always starts with the title and description of the topic. This was followed by the data of one of the eight variables and a description of the variable. Furthermore, all the statements of the discussion were attached to the current variable. The prompt instructed the LLM to create a summary based on the provided data. The most effective prompt was the one with "Group-aware consensus + Representativeness of Comments." (A.5.1). This prompt enabled the LLM to generate a summary that included the main statements of the topic. Moreover, it reflected the sentiments of the discussion participants. The summary was well-rounded. It highlighted the major points and concluded with an overview sentence.

In the second iteration, we extracted only the statements and group A and B voting results from the JSON graph object. The prompt for this iteration contained the topic, and description of the discussion, group A and B data, and an instruction to summarize. This approach successfully summarized the viewpoints of groups A and B in the discussion (A.5.2).

Afterward, the prompts of the paper "Opportunities and Risks of LLMs for Scalable Deliberation with Polis" (Small et al., 2023) were replicated. It was specifically A2 version 1, A2 version 2, and A7 version 2. A2 version 2 proved to be the most promising, as it accurately summarized the discussion and provided an overview of groups A and B. The only minor flaw was that it was written in a very machine-like manner, using terms such as "topic detected."

Lastly, for the prompt with raw data, the LLM was provided with the topic and description of the discussion, followed by the raw data (user votes and statements).

GPT was tasked to summarize the discussion. While the LLM could summarize the statements, it struggled to understand the sentiment of the discussion.

Overall, comparing all the different summaries generated by the LLM, it was found that the summary using "Group-aware consensus + Representativeness of Comments" combined with the summary of group A and B were the best. These summaries provided a detailed insight into the sentiment of the discussion and the perspectives of both groups. For a more detailed explanation, please refer to the A.5.1.

3.6.2. Prompt Tweaking

To create more efficient summaries, prompt tweaking was attempted, as explained in the previous section. The most promising prompt included the values of "Group-aware consensus + Representativeness of Comments." To identify the most effective prompt for generating graph summaries, three versions of the prompt were created. These versions varied in wording and detail in explaining the Group-aware consensus and Representativeness of Comments. Version 1 provided the most in-depth explanation of the variables, Version 2 offered the shortest explanation, and Version 3 presented a medium-length explanation.

During the test trials, it was found that Versions 1 and 2 effectively conveyed the key points regarding participants' attitudes towards the discussion (in this case, Unmanned Aerial Mobility). Yet, certain elements made one summary slightly better than the other. Version 1 was more comprehensive. The summary covered participants' willingness and interest in UAM and it dove deeper into specific concerns such as delivery services, noise pollution, and safety implications. This detailed perspective made Version 1 the better summary overall. In contrast, Version 3 often explained the variables instead of summarizing the discussion, which occurred multiple times.

Evaluating the different prompts demonstrated that Version 1 was the most effective for summarizing the polis graph. Therefore, this prompt was used in the main study A.5.1.

3.7. Prompt Ability to Chat

For the ability to chat, upon opening the polis page, the LLM generated a prompt to summarize the discussion. Furthermore, the user had the opportunity to ask questions directly in the chat. These questions were then sent to GPT 3.5-Turbo. The following prompt was utilized A.6. to summarize the discussion. The large language model was provided with the title, subtitle, and all comments. Additionally, the prompt was designed to summarize the main arguments of the discussion, identify common themes, and deliver a comprehensive overview that is easily understandable. Through multiple iterations, it was determined that this prompt effectively facilitates a proper understanding of the topic and discussion.

4. Hypothesis

Based on the Introduction, theoretical background knowledge, and prompt technique, the following hypothesis (H) was concluded: What are good summarization techniques for LLMs to sum up discussions on digital community deliberation tools? Hereby six different summarization techniques were evaluated through a within-subject design experiment.

First, it was assumed that the users of digital community deliberation tools would benefit from a summary that offers clickable citation links at the end. Furthermore, it was also evaluated whether the summary profited from being written in a simplified language. Moreover, it was explored if the users would like the summary to be always visible on the page or if they would instead initiate the summary themselves through clicking on a summary button. Also, the tree summary technique was evaluated, which had the ability to let users branch into different topics and dig deeper and deeper into specific areas of a topic. The digital community deliberation tool Polis offers a graph that summarizes the various viewpoints of the comments provided and user groups. Hereby, it was tried to summarise the graph's key characteristics, including user groups. Lastly, potential users were tested to see if they preferred the ability to ask the LLM questions about the summary through a chat.

H1:What are good summarization techniques for LLMs to sum up discussions on digital community deliberation tools?

5. Implementation

In order to test the hypothesis, a modified version of the polis platform was developed. Six discussions were created on the polis to test the six different summarisation techniques. During each technique, qualitative data was collected. After each discussion, the participants were presented with a questionnaire. With the questionnaire and the qualitative data analysis, it was tried to determine whether the presented technique had a positive or negative summary experience for the participants. Following, it will be described which parts of the polis platform were changed or added and how large language models were implemented into the platform.

5.1. UI Design

Firstly, in this section, the paper will talk about the implementation of the summary window. The summary window is the place where the user reads and interacts with the summary. Afterward, the UI changes for the independent variables participant initiated, tree summary and the ability to chat will be discussed.

5.1.1. Summary Window

All of the tested summary techniques were presented in a summary window on the page. As seen in [ABB] the current implementation of the polis platform offers blank white space on the left and right side of the page. For this experiment,the right-hand side was utilized to implement the summary window. The summary window uses the React-Chat-Window (Kingofthestack, n.d.) package. This package was modified according to the look and feel of the polis platform and the University of Texas at Austin. Therefore, all blue pieces were altered to burnt orange (BF5700), the color of the University of Texas at Austin. The size of the window was changed to fill the entire right side of the page. Moreover, it was also made sure that the window

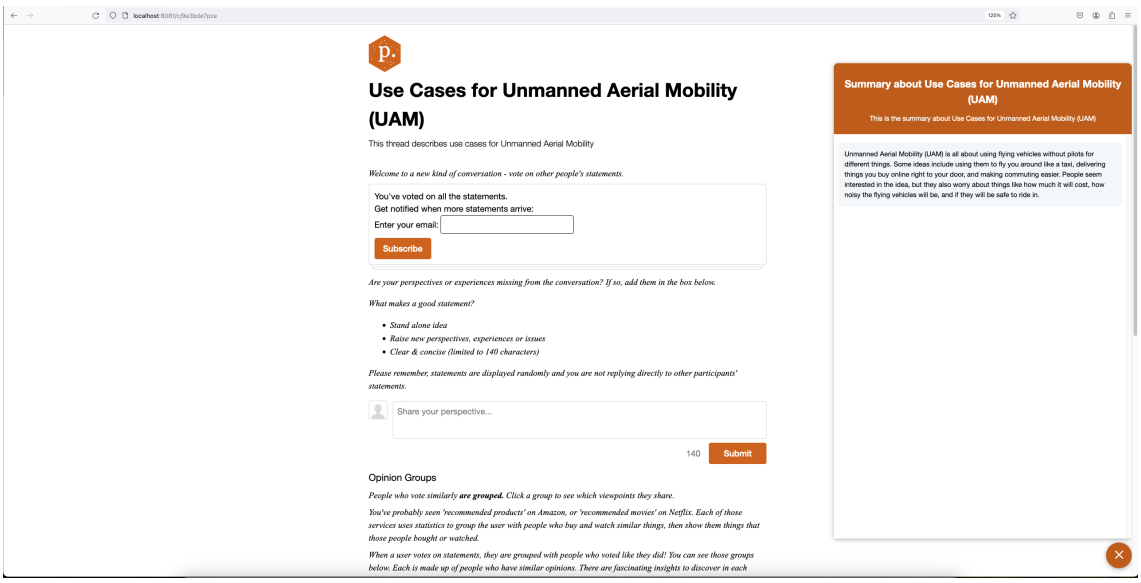

Figure 5.1.. Polis Summary Window

is not resizable. Furthermore, the summary title on the top was modified to spell "Summary about" and then the discussion title.

5.1.2. Participant Initiated

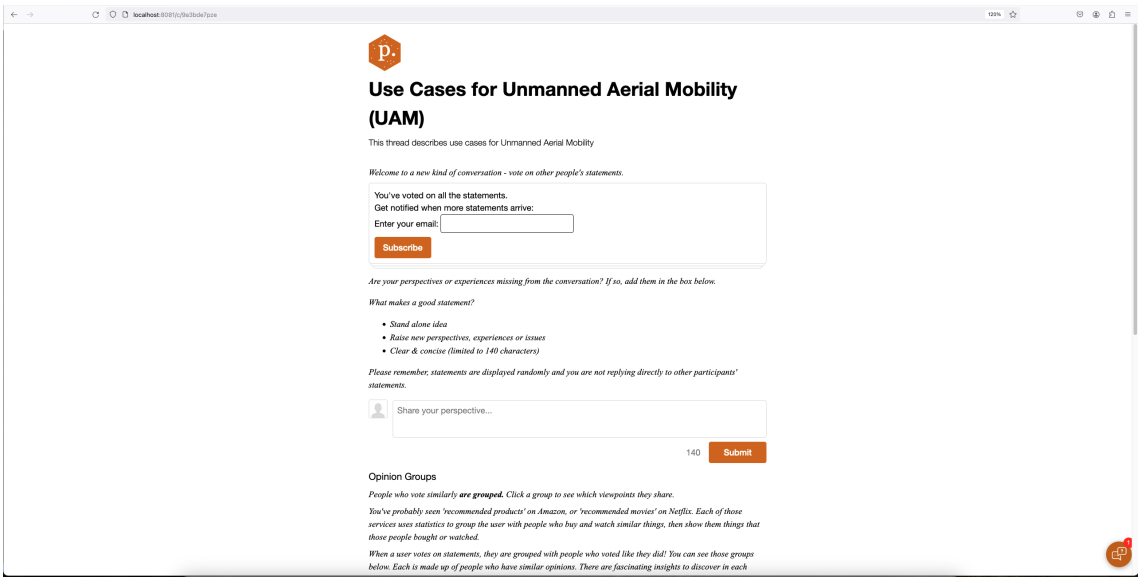

Figure 5.2.. Polis Participant Initiated

While testing this independent variable, the participants first had to click on the summary button (see [ABB]) to open the summary window. Hereby, the button was

placed in the bottom right corner, and it had a red notification bubble attached to it. The guidelines from the Norman and Nilsen Group (Budiu, 2018) were utilized for the decision to place the button in the bottom right corner. According to this article, if a button is placed in the bottom right corner, most users have the expectation, that it will be an interactive element. Furthermore, this position ensures that there will be no interference with the main task of the site, which is voting on the discussion. Due to that, it was decided to place the button in the bottom right corner.

The summary button has a red notification bubble attached to it in order to draw attention. According to the Apple Guidelines (Inc., 2024), this helps to get the user's attention and encourage engagement. During the experiment, it was necessary to engage with the summary. Therefore, a red notification bubble was attached to the summary button.

5.1.3. Tree Summary

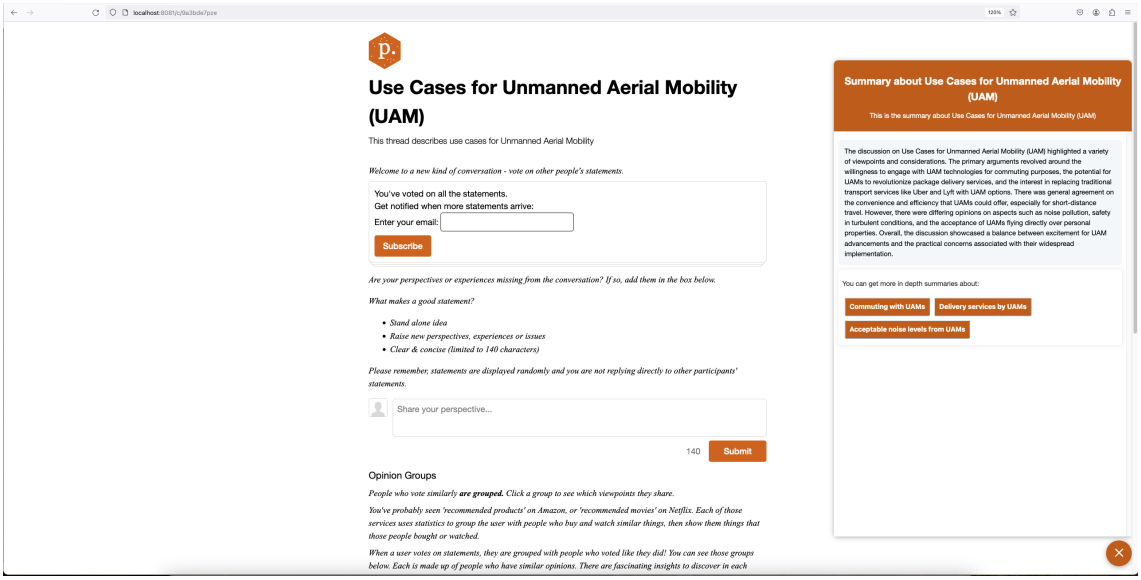

Figure 5.3.. Polis Tree Summary

The UI design for the tree summary was fundamentally inspired by the Wikium research paper (Zhang et al., 2017), particularly the layout of having a summary block underneath each other. Something that is different from Wikium's paper is that for this implementation, the user was first presented with a summary of the discussion. Underneath, three buttons were placed, which stated the main topics of the summary. If a user clicks on the button, a new summary will be created about

the clicked topic, which will appear at the end of the last summary. Moreover, from this summary, three new main topics will be identified and placed underneath the summary. The user could branch off with this system, and the participant could dig deeper into topics. This design of having a summary with three buttons was iteratively developed before the experiment started.

5.1.4. Ability to Chat

For the ability to chat, a text field was implemented where the user could type in questions that the LLM would answer. Hereby were chat elements of the React-Chat-Window (Kingofthestack, n.d.) package mainly used. The some of the elements were recolored to burnt orange to suit the look and feel of the webpage.

5.2. Prompt Implementation

The OpenAI API was utilized for the implementation of the prompts. The API was integrated following the recommendations in the documentation (OpenAI, 2024). Depending on the independent variable, one of the prompts discussed in Chapter 3 was sent to the API. The response was displayed in the summary window.

6. Methodology

6.1. Design

In order to understand what good summarization techniques are for LLMs to sum up discussions on digital community deliberation tools, an empirical study was conducted. This study used the digital community deliberation tool Polis ("Polis," 2024). Participants in this study were tasked with evaluating various summarization techniques implemented within Polis.

All participants evaluated all the six different summarization techniques. All of the techniques were presented to the participant in the following order: Summary with Citation, Simple Language Summary, Participant Initiated, Tree Structure, Graph Summary, and Chat Feature. The study is a within-subject design.

During the study, qualitative and quantitative data were collected. Qualitative data was collected through the Thinking-Aloud method. Therefore, the participant was prompted to verbalize his thoughts, feelings, and decisions as he explored the different summary techniques. Furthermore, potential follow-up questions were prepared to ask the users more in detail about the system. Quantitative data was obtained through a modified System Usability Scale (SUS) administered after each evaluation of the summarization technique. Participants rated their perceived usability of each technique on a Likert scale ranging from one to five.

A combination of qualitative and quantitative data was used to provide a comprehensive understanding of user experiences and preferences across different summarization techniques. This mixed-methods approach enabled a thorough evaluation of each technique's efficacy in enhancing user interaction and satisfaction within digital community deliberation tools.

6.2. Data Collection Methods

This study uses qualitative and quantitative data to evaluate what good summarization techniques are. For qualitative data, the Thinking-Aloud method was used. A modified System Usability Scale (SUS) was utilized to obtain quantitative data.

The Thinking-Aloud method (Ericsson & Simon, 1993) was used for cognitive transparency and to identify usability and interaction issues. The Thinking-Aloud method provides insights into the participant's thought processes, making it easier to understand how they interpret and make decisions based on the summaries provided by LLMs. Furthermore, it helps identify usability problems within the digital tool. This real-time feedback is crucial for identifying design and interaction issues that may not be apparent to designers or through quantitative data alone. During the study, the participants were asked to verbalize their thoughts, feelings, and decisions as they explored the different summary techniques. To further enhance the Thinking-Aloud method, a script with follow-up questions was implemented. For each of the six independent variables, the script offered four to five questions. The script would also offer further follow-up questions depending on the participant's answer. This data collection strategy led to rich qualitative data. After the study, all of the interviews were transcribed using otter.ai. Afterward, a codebook writeup was created from the transcribes to break down the most important statements of the participants.

For quantitative data analysis, a modified System Usability Scale (SUS) (Klug, 2017) was employed to evaluate the usability of the summarization techniques. The survey was utilized after each of the six independent variables. The SUS is a reliable tool for measuring user satisfaction and usability. In this case, a scale from 1 to 5 was utilized, where 1 indicates strong disagreement, and 5 indicates strong agreement. Participants responded to four specific questions that focused on various aspects of the summary's usability. These questions aimed to assess the ease of use, efficiency, and overall satisfaction with the summarization techniques. The quantitative data collected through the SUS provided a comprehensive metric of user experience, allowing for a systematic comparison of the different summarization techniques and their effectiveness in terms of user satisfaction and usability. This data was then analyzed to identify trends and draw conclusions. Nonetheless, it must be acknowledged that because of the small sample size of 14 participants, no p-test became

significant (no alpha =< .05).

6.3. Study Procedure

The study procedure consists of an introduction, warm-up, main task, and a closing part. In the introduction part, participants were introduced to the purpose of the study and the Think-Aloud method. Afterward, participants were asked to consent to recording the audio during the study.

After receiving the consent, the participant was introduced to the warm-up task. This section was implemented in order to train people on the Think-Aloud method because only some participants had experience with this technique. The warm-up task consisted of browsing through the Apple website (Apple, 2024) and looking through iPhone models.

At the beginning of the main task, the participant received a more in-depth explanation of the Polis platform and the layout of the website. Furthermore, the concept of unmanned area mobility was explained. Afterward, the user story was set in order to make the Thinking-Aloud easier. Then, for each of the six independent variables the study leader first read the task description. The task description explained that the participant should mainly interact with the summary and he should elaborate on his experience with it. Following, the participant interacted with the Polis website, which had one of the summary techniques implemented. For each iteration, another summary technique was tested. All of the summary techniques were presented in the following order: Summary with Citation, Simple Language Summary, Participant Initiated, Tree Structure, Graph Summary, and Chat Feature. During the interaction with the summary, ther users were asked questions from the Follow up Questions manuscript.

After each technique, the participants were asked to fill out a questionnaire. The questionnaire was an adjusted system usability survey. The survey used a scale from one to five. Hereby, one was labeled as strongly disagree, and five were labeled as strongly agree. At the end of the experiment, the purpose of the study was explained. Furthermore, the research leader thanked the participant for taking the time. In the end, the participant was given a chance to ask questions.

6.4. Experiment Sample

Fifteen individuals were recruited as participants for this study, with one individual designated for a pilot study and the remaining fourteen actively involved in the experimental phase. The cohort consited of students affiliated with the University of Texas at Austin, ensuring a homogeneous academic background within the sample.

Gender distribution within the participant pool was balanced, with six identifying as female and nine as male. Academic disciplines represented among participants encompassed a diverse spectrum. Six people have an engineering or a science-related major. Five people have a liberal arts background. Furthermore, there was a student in Communication, Law, Journalism, and Business, respectively.

Regarding prior knowledge and experience of the study's focus, it is noteworthy that only one participant possessed prior familiarity with unmanned area mobility. Five participants reported regular utilization of LLMs. Geographical diversity was also accounted for, with ten participants from the United States and five from Europe. All European participants, despite being nonnative English speakers, demonstrated language proficiency at the C1 level.

Throughout the experimental protocol, all participants systematically evaluated six distinct summary techniques. They all enriched the study with their comprehensive feedback.

7. Result

In this section, the results of the study will be discussed. Hereby, we will talk about how the participants generally perceived the summary. Following, the results of each of the six independent variables will be explored. Afterward, the differences between participants from the US and those from Europe will be discussed in detail. At the end of this section, the paper is going to elaborate on how the different college majors of the participants might have impacted the perceptions of the summary techniques.

7.1. Adding Summary in General

Almost all participants expressed that they would prefer it if the polis platform offered a summary. Participant 1 emphasized positively that if she "had summaries like this [she would] feel like [she] would engage more". Participants have also claimed that the summary would help them to get a fast understanding of the discussion.

They also have pointed out general opportunities for improvement. Participants three, seven, and eleven expressed that they would favor a different writing style for the summary. They would prefer to have bullet points instead of plain text. They claim that bullet points would make it faster to understand the essential points of the summary. Another area of improvement is the summary window itself. It was noted that the summary window would visually appear similar to a FAQ or a chat assistant (P3, P10). The participants describe that they have encountered numerous websites where the FAQ or chat assistant is placed on the right side of the page, and the window structure looks similar. Because of that, users might mistake the summary window for something else through top-down knowledge. Moreover, it was mentioned that the summary sounded biased. The summary would stress more the positive aspects than the negative. Participant 10, for example, stated: "I feel like this kind of reads to me like it's like a marketing blurb or like something you put in

like pitch like a venture capitalist.". Participants who mentioned that the summary might be biased have expressed that it negatively impacted their opinion about the summary.

During the experiment, participants mentioned that they most fancy the citation and tree structure. Furthermore, participants 11 and 13 have expressed that they would like both techniques combined, the tree structure and the citation feature.

7.2. Summary with Citation

Generally, people fancied summaries, which had citations. The second participant said that the citation helped her understand unmanned area mobility. Furthermore, it was described that by clicking on the citation and reading further into the topic, the user could gain new perspectives (P4). Providing citations conveys to users that the summary is trustworthy (P7). This could be something positive because it enhances the credibility of the information presented. Additionally, citations offer users the opportunity to explore the original sources, deepening their understanding of the topic.

According to the survey, the summary with citations was intuitive. Participants disagreed with the idea of needing the support of a technical person to be able to use this summary system (mean = 1.62).

It could also be negative because LLMs can sometimes hallucinate, generating information that appears plausible but is inaccurate or entirely fabricated. If participants do not crosscheck the summaries provided by the LLMs, they might accept these hallucinations as truths. When asked how often they would click on the citations, some participants explained that they would only click on the citation link if they were interested in the topic, if they didn't understand it, or when they would question the information provided by the summary. For example, participant 1 stated: "If I would be really interested in the topic, I would for sure click on the citations. However, if it is something really simple, then I would just go with the summary provided".

Overall, participants generally appreciated citations for their trustworthiness and utility. Whether participants would click on the citation would depend on their interest and need for clarification on the topic.

7.3. Simple Language Summary

Summaries that used simple language received mixed evaluations. On the one hand side, users fancied the simple language and emphasized that it would be easier to understand. On the other hand side, some participants found this type of summary not sufficient to explain the topic and to be an introduction rather than a summary (P13).

The ones who emphasized the simple language would express that: “It’s easier to understand especially for people who maybe know who like, straightforward language.” (P3). Because it is easier to understand, it makes it also more suitable for people with a lower level of education. Moreover, simple language allows foreigners to grasp the discussion more easily (P8). Including more foreigners and individuals with lower education levels would enhance accessibility.

Some participants have pointed out that there is room for improvement. Participant 7 would like the summary to go more in-depth. He stated that, according to the current summary, he hasn’t gained a sufficient understanding of unmanned area mobility and the discussion. Moreover, the participants described that the simple language would use words, which makes it seem odd. The participants described that they felt an uncanny valley phenomenon. The summary would use words like “cool” which made the summary sound unsuitable.

In conclusion, some users appreciated the clarity and accessibility, while others felt these summaries lacked depth and were more introductory than informative.

7.4. Participant Initiated

Participants had varied reactions to the participant-initiated summary feature. The notification design and placement were generally well-received. Participant 1 noted that the red notification bubble made it clear that a new message was available, effectively signaling additional information. This view was also expressed by Participant 2, who appreciated the red notification bubble and felt it clearly indicated the summary’s presence.

However, some participants suggested improvements. Participant 7 highlighted the need for better visibility of the summary button upon page load to prevent it from being mistaken for an FAQ or chat assistant icon. Similarly, Participant 8 recommended making the button larger and more distinguishable to avoid confusion with other elements like chatbots. Overall, participants valued visual cues and clear labeling to enhance usability. For instance, Participant 2 proposed that hovering over the icon could trigger a tooltip explaining its function as a summary. Participant 10 suggested including the GPT logo to signify the summary is AI-generated. This would leverage familiarity with existing AI tools.

The participant-initiated feature was found useful and intuitive, but improvements such as larger icons, clearer labels, and additional visual cues were suggested. This would enhance the user experience and ensure easy recognition and access.

7.5. Tree Summary

The tree structure summary feature received positive feedback from participants. They fancy the organization and interactive nature. Participant 3 found the tree structure straightforward and helpful. He stated that it allowed users to access the exact information they wanted without having to skim through the entire text. Similarly, Participant 12 favored the tree structure over reading long research papers. He claimed it would provide a more efficient way to navigate through information. He emphasized that the buttons made the feature very intuitive, making it easy for him to interact with the system. Furthermore, this intuitivity is also shown in the survey. According to the survey, participants agree that most people would learn to use this summary system very quickly (mean = 4.23). Moreover, Participant 11 compared the tree structure to a "smart FAQ," highlighting its ability to offer more depth and detailed exploration of topics.

However, participants also suggested opportunities for improvements. Some participants struggled with keeping track of the different branches of summaries. For example, Participant 7 suggested the inclusion of a feature to close the summary after reading and the ability to jump between different sections of the tree. He mentioned that this could help manage the expanding nature of the tree structure. This would allow users to navigate more efficiently.

Participants also noted the importance of visual cues and clear labeling. For instance, Participant 10 proposed the use of visual signifiers similar to those used in GitHub, to indicate branching points within the tree. This could help users understand the structure at a glance and make navigation more intuitive. Furthermore, during the experiment, multiple participants suggested that the tree structure could be merged together with the citation feature. For example, participant 8 wished that the tree structure would be merged with other features for a more seamless experience. Participant 13 loved the ability to dive deeper into topics using the tree structure. He also expressed that he would like sources to be included within this format. Participants explained that citations would add credibility to the summaries provided.

In conclusion, the tree structure summary feature was found to be highly beneficial for its depth and interactivity. Many participants recommended enhancements such as merging with the citation feature. Also, it was mentioned that visual cues should be included for an easier-to-understand structure. These changes could further enhance usability and ensure that users can easily navigate and extract valuable information from the summaries.

7.6. Graph Summary

Participants provided a variety of feedback on the graph summary feature. Generally, they appreciated the explanations of the graphs. However, some participants found the graphs themselves challenging to interpret. The missing labels and unconventional shapes were described as challenging by Participant 1.

During the study, opportunities for improvement were identified. For example, participant 3 suggested that the initial text explaining the graph could be shorter, as it did not significantly help his understanding of the visual data. Furthermore, Participant 10 mentioned that the information provided was overwhelming and suggested a visualization. He expressed the idea that the LLM could generate a graph with nodes to make the data more comprehensible. Therefore, he stated: “There’s like a central node that you can make, and then you have ideas branching off of that.”. He mentioned that the graph can be interactable like the program Obsidian.

Participants also expressed the need for clearer labeling and more contrast between different groups in the graphs. Participant 8 emphasized the need for a clear distinction between group A and group B to avoid confusion. Additionally, participants indicated a preference for concise summaries that capture the key points of the graph data. Participant 11 suggested that bullet points would make the summary easier to digest and quicker to understand. The desire for a more interactive and dynamic summary experience was echoed by Participant 12, who wanted the ability to replace old summaries with new ones seamlessly. He appreciated the detailed explanations of group differences but felt that a more streamlined presentation could improve accessibility.

Overall, while participants found the graph summaries helpful, they recommended enhancements such as better labeling, clearer distinctions between groups A and B, and more interactive features to improve usability and comprehension. Incorporating these suggestions could make graph summaries more effective in conveying complex information clearly and efficiently.

7.7. Ability to Chat

The chat feature received positive feedback from most participants, highlighting its utility and user-friendly design. Participant 1 stated, “I really think it was a good response,”. Hereby, she indicated her satisfaction with the chat’s ability to provide important information. Similarly, Participant 2 expressed a preference for using the chat function frequently, particularly when interested in a topic. However, there were concerns about the reliability of the information provided by the chat. It was suggested that a disclaimer should be included to inform users that the chat might not always provide accurate information, similar to disclaimers used by ChatGPT (P3).

The design and interface of the chat feature were generally well-received. Participant 4 mentioned that the chat’s UI was intuitive and resembled typical chat applications, which made it easy to use. In the survey, participants also strongly agreed that this summary system would be very quick to learn (4.38). Participant 10 also confirmed that, the familiar design enhanced their willingness to interact with the chat feature. Despite these positive affirmations, some participants suggested improvements to enhance clarity and usability. For example, Participant 7 recommended adding the

GPT logo to indicate that the responses were AI-generated. This would leverage the existing recognition of AI tools to boost user trust.

Overall, while participants appreciated the chat feature for its accessibility and ease of use, they also highlighted the need for clear disclaimers about the accuracy of the information and suggested incorporating familiar AI branding to enhance trust and usability. These improvements could make the chat feature a more reliable and engaging tool for users seeking quick and interactive information.

7.8. Results based on University Major

The sample size for the different college majors is quite small. Therefore, the following results should be interpreted as indicative trends rather than definitive conclusions. Out of six participants from STEM or Engineering backgrounds, four expressed a preference for the tree summary. These participants highlighted their appreciation for its interactivity and depth. Conversely, five out of six participants from Liberal Arts or Communication fields favored the chat feature. They described enjoying the system's ease of use. These observations suggest that preferences for summary techniques may vary based on the participant's college major or area of interest.

7.9. Results based on Nationality

During the interviews, most participant's statements showed no significant correlation between their nationality and the nature of their comments. However, a notable finding was that concerns about LLMs generating incorrect answers or hallucinations were predominantly raised by European participants. In this study, 10 participants were from the US and 5 from Europe. 1 US participant and 4 European participants expressed this matter. The participants suggested incorporating a disclaimer in the summary window to alert users to the possibility of LLM hallucinations. This could mean that the presentation of LLM summaries might need to be tailored differently for different countries to address these concerns effectively.

8. Discussion

The study investigated the effectiveness of various summarization techniques using LLMs to enhance online community deliberation. Six different summarization methods were investigated: summaries with citations, simple language summaries, participant initiated summaries, tree-structured summaries, graph summaries, and summaries with the ability to chat. The primary aim was to identify the techniques that most effectively aid users in understanding and engaging with deliberative content.

8.1. Summary with Citation

Participants generally valued the summary with citations. They claimed that it added an element of trustworthiness and allowed for deeper exploration of topics. However, there is a potential risk of overreliance on the cited information, especially given the possibility of LLM hallucinations. Furthermore, users indicated that they would click on citations primarily when they were particularly interested in the topic or required clarification.

8.2. Simple Language Summary

The simple language summaries received mixed feedback. On the one hand side, some participants appreciated it for their clarity and accessibility, particularly for non-native English speakers and individuals with lower educational backgrounds. On the other hand, some participants felt that the simplicity resulted in a lack of depth and a more introductory rather than broad understanding of the discussion.

8.3. Participant Initiated

The participant-initiated summary feature was received with mixed sentiment. Some participants found the participant-initiated approach utterly intuitive. However, improvements were suggested to enhance visibility and avoid confusion with other interface elements like chatbots. Moreover, users valued clear labeling and visual cues that indicated the presence of additional information.

8.4. Tree Structure

The tree-structured summary was highly favored for its interactive nature and the ability to dive deeper into topics. Participants appreciated the organization and depth of this method. However, suggestions for improvements included better navigation or visualization of the different summaries was suggested.

8.5. Graph Summary

The graph summaries provided valuable insights. It also helped participants understand the graph. Furthermore, it helped to inform the user about the sentiment of the discussion. Yet, participants recommended a clearer distinction between groups A and B.

8.6. Ability to Chat

The ability to chat was positively received for its user-friendly design and the ability to provide immediate information. However, concerns were raised about the reliability of the information, and participants suggested including disclaimers about potential misinformation. The familiar interface of chat applications made this feature particularly intuitive and engaging.

8.7. Limitation

The study's sample size was relatively small, and further research with a larger and more diverse participant pool is necessary to generalize the findings.

9. Future work

Overall, this experiment was successful. Each of the six independent variables demonstrated advantages and potential for further investigation.

9.1. Summary with Citation

Results from this study have shown that citations enhance the perceived trustworthiness of a summary. Future research could explore this further, particularly the potential issues arising when users overtrust a summary simply because it includes citations. This is especially relevant given that some participants mentioned they rarely verify citations. Investigating the conditions under which users trust summaries and the potential risks of unverified information could provide valuable insights.

9.2. Simple Langauge Summary

Further studies could explore the effectiveness of summaries with different age groups. For instance, they could investigate how participants respond to summaries, which are prompted for 14, 16, and 18-year-olds, and compare the results. Additionally, future research could explore the optimal length for a summary.

9.3. Participant Initiated

Future research could explore the placement of the summary button within the main content of the page. Therefore, it will be placed more prominently in the middle of the page. Furthermore, different button icons could be investigated to find out which icons best signal LLM summary. By experimenting with these design

changes, future researchers could investigate whether a more prominently placed and visually appealing summary button icon encourages greater user engagement with the summary.

9.4. Tree Structure

Future researchers can further develop the tree structure of the summaries. Building on the approaches from this paper, they could enhance the user interface to make the tree structure more organized and user-friendly. As suggested by Participant 10, implementing a git history-like structure could be beneficial for following the different branches of the summary, providing users with a clear and intuitive way to navigate through the information.

9.5. Graph Summary

While the graph summary provided valuable insights, there is room for improvement in directly comparing Group A and Group B. Future research could explore the development of a prompt that facilitates direct comparisons between both groups. This approach would make it clearer for users to understand the relative standings of each group, enhancing the utility and comprehensibility of the graph summary.

9.6. Ability to Chat

Future research could conduct an A/B test to evaluate the impact of disclaimers that the LLM could produce wrong information. In one scenario, users would have the ability to ask questions through the chat with a disclaimer indicating that LLM could possibly hallucinate. The chat would be available without such a disclaimer in the other scenario. This study could investigate how the presence or absence of the disclaimer affects user behavior and their approach to asking questions in the chat.

9.7. Merge Tree Structure and Summary with Citation

Three participants suggested integrating the summary with citations within the tree summary structure. Future research could test this combined approach to determine if it offers a better user experience. By merging these two methods, researchers could evaluate whether the tree structure and added citations improve comprehension and trustworthiness.

Dieses Dokument hat auch einen Anhang A ab Seite 45.

Bibliography

Agency, E. U. A. S. (2024). Drones and air mobility landscape: Basics explained [Accessed: 2024-06-03]. (Cit. on p. 7).

American Academy of Child Adolescent Psychiatry. (2016). The teen brain: Behavior, problem solving, and decision making [Accessed: 2024-05-30]. (Cit. on p. 10).

Apple. (2024). Apple [Accessed: June 5, 2024]. (Cit. on p. 25).

Bobbio, L. (2019). Designing effective public participation. *Policy and Society*, *38*(1), 41–57 (cit. on p. 2).

Britannica, E. (2020). Ancient rome [Accessed: 2024-06-03]. (Cit. on p. 1).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901 (cit. on p. 7).

Budiu, R. (2018). The user experience of customer-service chat: 20 guidelines [Accessed: 2024-05-28]. *Nielsen Norman Group* (cit. on p. 21).

Ericsson, K., & Simon, H. (1993). Protocol analysis: Verbal reports as data mit press. cambridge, ma.[links] flavell, jh (1976). metacognitive aspects of problem solving. *The Nature of Intelligence. Hills-dale. NJ: Erlbaum* (cit. on p. 24).

Inc., A. (2024). Notifications - human interface guidelines [Accessed: 2024-05-28]. (Cit. on p. 21).

Jackson. (2022, April). Confused user experience with polis – why can't i explore groups in more detail? [GitHub Discussion #1368]. (Cit. on p. 12).

Kingofthestack. (n.d.). React-chat-window [Accessed: 2024-05-21]. (Cit. on pp. 19, 22).

Klug, B. (2017). An overview of the system usability scale in library website and system usability testing. *Weave: Journal of Library User Experience*, *1*(6) (cit. on p. 24).

Kriplean, T., Morgan, J., Freelon, D., Borning, A., & Bennett, L. (2012). Supporting reflective public thought with considerit. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 265–274 (cit. on p. 2).

Lark. (2024). One-shot learning [Accessed: 2024-06-07]. (Cit. on p. 7).

Levine, P., Fung, A., & Gastil, J. (2005). Future directions for public deliberation. *Journal of Deliberative Democracy*, *1*(1) (cit. on p. 1).

OpenAI. (2024). Openai platform quickstart documentation. (Cit. on p. 22).

Polis [Accessed: June 5, 2024]. (2024). https://pol.is/home (cit. on pp. 5, 23).

Robertson, N. G. (2023). Ecclesia - assembly of athens [Accessed: 2024-06-03]. (Cit. on p. 1).

Shortall, R., Itten, A., van der Meer, M., Murukannaiah, P., & Jonker, C. (2022). Reason against the machine? future directions for mass online deliberation. *Frontiers in Political Science*, *4* (cit. on p. 2).

Small, C. T., Vendrov, I., Durmus, E., Homaei, H., Barry, E., Cornebise, J., Suzman, T., Ganguli, D., & Megill, C. (2023). Opportunities and risks of llms for scalable deliberation with polis. *arXiv preprint arXiv:2306.11932* (cit. on pp. 2, 12 sq.).

Staff, N. (2024). Participation at scale can repair the public square [Accessed: 2024-06-03]. *Noema* (cit. on p. 6).

University of Houston-Downtown Center for Public Deliberation. (2024). What is public deliberation? [Accessed: 2024-06-03]. (Cit. on p. 1).

Watanangura, P., Vanichrudee, S., Minteer, O., Sringamdee, T., Thanngam, N., & Siriborvornratanakul, T. (2023). A comparative survey of text summarization techniques. *SN Computer Science*, *5*(1), 47 (cit. on p. 9).

Yeo, S., Lim, G., Gao, J., Zhang, W., & Perrault, S. T. (2024). Help me reflect: Leveraging self-reflection interface nudges to enhance deliberativeness on online deliberation platforms. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–32 (cit. on p. 2).

Zhang, A. X., Verou, L., & Karger, D. (2017). Wikum: Bridging discussion forums and wikis using recursive summarization. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2082–2096 (cit. on pp. 2, 11, 21).

Appendix A. Prompts

A.1. Summary with Citation

const citationSummaryPrompt = "Please summarize" + props.response.conversation.topic + props.response.conversation.description + ". Important: Please provide links to scientific resources about the information you will give me. Please highlight the links through markup annotation."

A.2. Simple Language Summary

const SimpleLanguageSummaryPrompt = "Please summarize" + props.response.conversation.topic + props.response.conversation.description + ". Please explain in simple language, don't use abstract terms adn explain as if I was a " + "16-year-old" + "person."

A.3. Participant Initiated

const CommentSummaryPrompt = "Please summarize the discussion" + props.response.conversation.to + "These are the discussion statements" + discussionstatements + props.response.conversation.descripti + "Determine the primary arguments or viewpoints from the discussion." + "Identify any common themes or points of agreement among the comments." + "This summary will offer a comprehensive overview of the discussion, enabling readers to quickly understand the key topics and the spectrum of views presented."

A.4. Tree Summary

A.4.1. Discussion Summarization

const TreeCommentSummaryPrompt = "Please summarize the discussion" + props.response.convers
+ "These are the discussion statements" + discussionstatements + props.response.conversation.descr
+ "Determine the primary arguments or viewpoints from the discussion." + "Identify any common themes or points of agreement among the comments." + "This summary will offer a comprehensive overview of the discussion, enabling readers to quickly understand the key topics and the spectrum of views presented."

A.4.2. Main Topics Prompt

const MainTopicsPrompt = "This is the discussion" + props.response.conversation.topic
+ "These are the discussion statements" + discussionstatements + props.response.conversation.descr
+ "Please identify the three main topics of the discussion. Please provide them in the following schema [topic1] [topic2] [topic3]."

A.4.3. Further Summary

const FurtherSummaryPrompt = "Please summarize the discussion" + props.response.conversation.t
+ "This is the main topic of discussion: " + buttonName + "Determine the primary arguments or viewpoints from the discussion." + "Identify any common themes or points of agreement among the comments." + "This summary will offer a comprehensive overview of the discussion, enabling readers to quickly understand the key topics and the spectrum of views presented."

A.4.4. Further Main Topics Prompt

const FurtherMainTopicsPrompt = "Please summarize the discussion" + props.response.conversatio
+ "This is the main topic of discussion: " + buttonName + "Determine the primary arguments or viewpoints from the discussion." + "Identify any common themes or

points of agreement among the comments." + "This summary will offer a comprehensive overview of the discussion, enabling readers to quickly understand the key topics and the spectrum of views presented."

A.5. Graph Summary

A.5.1. Group-aware consensus and Representativeness of Comments

data = return txt: item?.txt, tid: item?.tid, bestAgree: repnessData['best-agree'], nAgree: repnessData['n-agree'], nSuccess: repnessData['n-success'], nTrials: repnessData['n-trials'], pSuccess: repnessData['p-success'], pTest: repnessData['p-test'], repfulFor: repnessData['repful-for'], repness: repnessData.repness, repnessTest: repnessData['repness-test'], groupAwareConsensus: consensus, ;

const GraphSummarization = "This is the topic of the conversation:" + props.response.conversation.topi
+ "The graph in the discussion was made by voting on the following statements with agree, disagree or skip on statements." + data + + "Every statement has a group aware consensus. This is a measure of the extent to which an opinion group in the conversation agrees (by vote) in response to a particular comment. " + + "Furthermore, some statement have a bestAgree value, which describes the highest level of agreement or consensus within a specific dataset or group." + + "Some statements have a nAgree value, this is the number of entities or participants who have agreed or provided a positive response in a given context. It quantifies agreement."+ + "Some statements have a nSuccess value, this counts the number of successful outcomes or instances within a set of trials or attempts."+ + "Some statements have a nTrials value, This represents the total number of trials, tests, or attempts conducted."+ + "Some statements have a pSuccess value, this is the probability of success or the success rate. It is calculated based on the ratio of successful outcomes to the total number of trials or attempts."+ + "Some statements have a repfulFor value, this indicate a measure or a feature that represents how representative or reputable something is"+ + "Some statements have a repness value, this quantifies how well a sample or a set of data represents a larger population or set of characteristics. "+ + "Some statements have a repnessTest value, This is a test or evaluation related to the 'repness' measure. "+ "Please analyze the data to understand what the values indicate about the comments. Interpret whether the participants agree with the

statements and describe the overall mood. Please give information on what key points. Make it sound like a human would say it."

A.5.2. Group A and B Summary

groupA = return txt: item?.txt, tid: item?.tid, Agree: voteStats.A, Disagree: voteStats.D, Skip: voteStats.S, ;

groupB = return txt: item?.txt, tid: item?.tid, Agree: voteStats.A, Disagree: voteStats.D, Skip: voteStats.S, ;

const CommentSummaryPromptA = "The discussion is about:" + props.response.conversation.topic
+ "This is an object with the discussion statements and how many people of group A agree, disagree or skipped the statement" + groupA + "Please analyze the data to understand what the values indicate about the comments. Interpret whether the participants of group A agree with the statements and describe the overall mood of group A. Please give information on what key points the group A agrees on and what key points the group A disagrees on. Make it sound like a human would say it."

const CommentSummaryPromptB = "The discussion is about:" + props.response.conversation.topic
+ "This is an object with the discussion statements and how many people of group B agree, disagree or skipped the statement" + groupB + "Please analyze the data to understand what the values indicate about the comments. Interpret whether the participants of group B agree with the statements and describe the overall mood of group B. Please give information on what key points the group B agrees on and what key points the group B disagrees on. Make it sound like a human would say it."

A.6. Prompt Ability to Chat

const PromptAbilitytoChat = "Please summarize the discussion" + props.response.conversation.topi
+ "These are the discussion statements" + discussionstatements + props.response.conversation.descr
+ "Determine the primary arguments or viewpoints from the discussion." + "Identify any common themes or points of agreement among the comments." + "This

summary will offer a comprehensive overview of the discussion, enabling readers to quickly understand the key topics and the spectrum of views presented."

High-Level Participant Overview Notes: *Notes by each participant, alongside some memorable quotes from their interview.*

Pilot 1: Male, White, phd researcher
Requested Design changes:

- “maybe no resizing” -> the summary window shouldn’t be resizable -> issue fixed
- “ui wise I would want the summary to be more central”
- “I would also like to know why it was generated in the middle of the night” -> In the summarization window, it seemed (through a small time display in the left bottom corner) that the time was 3:11 a.m. when the summary was generated -> That was not the case -> the time stamp was removed afterwards.

Requested Summary changes:

- “I would like to know why this difference between the two summaries exist” -> in this particular case there was no difference, afterwards the study design was changed so no for every discussion a new independent variable gets tested.
- Around 42 minutes the participants that that it would be hard for the “you can’t just read the summary and assume it’s a truth machine” -> hallucinations | “intricate power dynamics that play” -> summary is about the discussion comments not about a subjective view
- Around 55 minutes: “Ahh so the text was different but the system was the same”

Requested Interview changes:

- “maybe clarify what is being summarized” -> the new interview scripts talk in more clarity about summarizing the discussion comments.

P1: Female, White, STEM undergraduate
Citation:

- Around min 7 “If I had summaries like this I feel like I would engage more” -> she talked about that she prefers a summary in contrast to no summary
- Around min 8 “If I would be really interested in the topic I would for sure click on the citations however if it is something really simple then I would just go with the summary provided” -> she doesn’t know about UAM therefore a summary is helpful for her

Simple language:

- Around min 11 “I think it was the perfect use of language”
- Around min 11 There is no like overly complicated translation between sentences and no overly complicated terminology being used -> The participant is not a native speaker maybe therefore it is easier for her to understand the topic
- Around min 12 “Maybe instead of just one paragraph divide it into two or three so it is easier to read.”

Participant initiated:

- Due to the design the participant said it was clear to her that she would receive a message. Especially the red 1 made it clear to her

Tree Summary:

- Around 16 min “The fact that there were buttons now it just made me feel like I needed to click them”
- Around 17 min The participant rather wants bullet points instead of the summary

- Around 17 min “It is nice to have options and know that might not know everything about this specific topic”
- Participant asked me if there is any possibility to type in questions to ask more specific

Graph Summary:

- Around min 20 Participant about the graph: “I feel like this shapes make it a little weird to understand and there are no labeled axis to understand
- Around min 21 It [talking about the summary] explained very in detail what the graph was about
- Around min 21 The participant like that overview group A group B structure
- Around min 22 Especially she liked that it said response from group A, responses from group B -> this beginning string gets automatically added
- Around min 24 I would like a very short paragraphs max 2 sentences and then I would like the option to elaborate more by clicking on a button -> she is talking about the tree structure

Chat feature:

- Around min 25 “I really think it was a good response “ She likes the chat
- Around min 27 “I would use the chat function often if it was an interested in the topic”
- Actually I am not 100 % sure, chat gpt has a disclaimer that it is not always right, so it would be nice to have a disclaimer on this chat as well

P2: Female, White, Liberal Arts

Citation:

- At 5:13 min “there's a lot of text and maybe break it up into, like separate text bubbles, so it's easier to read.”
- At 5:13 min “The information is helpful, like regarding what the UAM “

Simple language:

- At 8:36 min “No” -> she sees no improvement in the language
- At 9:02 min “ I didn't know I wouldn't think to think about this about the environment before reading the summary”

Participant initiated:

- At 10:45 min “if you hover over it, then it would like pop up something that says like, summary” -> the user wants a tool tip
- At 11:27 min the user likes the red notification bubble
- At 11:27 min “the icon that kind of looks like Google Translate. So it seems like it would translate the page.”

Graph Summary:

- At 13:32 min “It described each group really well but also used to simple language.”, Furthermore: “ easy to understand vocabulary. So, it was easy to pick up on what the groups meant.”
- At 14:20 min I asked if she would improve something, she answered: “No, I think it does a good job.”

Tree summary:

- At 15:23 min “Their response time is kind of lagging” she clicked on a button and it took like 0.5 sec to generate a new summary
- At 15:57 min “ Do you feel because you can get more information that you also gain a deeper understanding? Yes.”

Chat feature:

- At 18:00 min she would use it to explain concepts
- At 18:20 min she likes the ui of the chat feature: “think it looks like a typical chat like how this is at the bottom, the text boxes at the bottom”

P3: Male, White, Liberal Arts
Citation:

- At 4:30 min “I really like is that, the summary points out the advantages and disadvantages of UAM”
- At 5:50 min “I would use it if I want to, let's say if I'm, yeah, if I'm really interested in this topic”
- At 6:10 min he describes when he would use the summary “It depends on on how how your interest level is and how much time you have, you know, if you see this, an online newspaper, you're just reading through and you've got time you're like, interested in the topic?”
- At 6:10 min “I think people will click on that.”
- At 6:52 min “Yeah, I think it's, it's really good. I also like the it's like at the very end so people can decide whether they want to engage more with this topic or not.”

Simple language:

- At 9:21 min “Like it's more it's like easier language. It's easier to understand especially for people who maybe know who who like like, straightforward language. I think they would like that one better.”
- At 9:21 min “it reminds me more of like, something you could find in a museum.”
- At 10:19 criticism “ I got the same information out of this as the summary before, maybe a little less, actually, because it's like a little shorter, and more like, straightforward.”
- At 12:06 “It has some, like some some some cool words and like flying machines. That sounds exciting. So I think it's um, yeah, it makes people excited about the topic more than the textbook.”

Participant initiated:

- At 14:25 min “It's definitely not the first thing that jumped into my motivation.”
- At 14:49 min “visual effects that make it more more obvious because it's like in the bottom right corner. So it's like, it's not not like the center of attention isn't my center of attention was like, in the middle.”
- At 15:52 min “It has like an icon that hints at the summary.”

Tree structure:

- At 18:15 min “Yes, I think that that's very straightforward, because you have to you have the, the inflammation right here”
- At 20:09 min “ Yeah, you can you can you can. You can choose what, what kind of information you're interested in. And your original website only had like, you're talking about the very first, you're the very first. Okay? Yeah. Yeah, definitely. Because you can you can you can basically, you can choose what you're most interested in.”
- At 20:47 min “And it's easy to use because it's just have to click the button and you get like, the exact information what you want in detail”
- At 20:47 min “And you don't have to like, I don't know, skim, skim through the whole text, you can like, get to your information right away. It's very, very helpful tool.”

Graph Summary:

- At 25:02 min “think it's like the depart for group A and B is as good as a good detail.”
- At 25:02 min “ the first part could be a little shorter. Because it didn't help me that much to understand the graphic, maybe maybe we can shorten it. The first part.”

Chat feature:

- At 27:54 min “ would have used the tool maybe to like, if it's let's say, I want to learn more about methods, advantages, disadvantages, like given bullet points?”
- At 28:49 min he says he thinks it is intuitive “Because it reminded me a little bit of objectivity. Yeah, because I used I use Chat GPT”

P4: female, white, liberal arts

Citation:

- At 4:44 min “ I think the summary, It helps, it really highlights the points about it and what they're talking about. Otherwise, I think I would just skip over whatever the conversation was.”
- At 5:35 min how often she would click on the citation: “maybe like, 10% of the time, if they were saying something really outlandish, like, you UAM's are going to give us alien content, you know, like, something like that? I'd be like, what, where? Why? You know what I mean? But if it's just like this, like general information, I would be less likely to click on the sources, would”
- At 6:04 min talking about the citation style and how she would change it: “If it had like a catchy title? I think that's always like, like the hook. It's kind of like clickbait”
- At 7:03 min she talks that summaries will give her new perspectives that she hadn’t thought of prior.

Simple language:

- At 8:50 min she realized by herself that this is more of a simple language: “ I feel like this one is a lot more simple. Like, it's definitely more maybe accessible to people who were like, oh, big science language, like I don't want to think about I don't want to read this because it looks like it has big words in it.”

- At 8:50 min she mentioned that probably simple language is better then citation for a broader audience: “ do notice that there's not any more citations, which also, not everybody likes citations or cares about citations. So I don't think that's necessarily a bad thing. ”
- At 9:27 min she says that the simple language is key in her eyes and that she would engage with it more often: “I think I'd be more inclined to read it.”, “I think it's more accessible for sure.”
- At 10:48 min in general for the summaries she wants “stylistically, like bolding kind of important words”

Participant initiated:

- At 12:19 min “maybe more at the top here somewhere somewhere else”
- At 12:58 mi “It reminds me of one of those like chat boxes, like where you like go to ask questions. So like, Yeah, I think that people would know, to click on it, or at least assume that that's what it is.”
- At 13:33 min “Because notifications bother me. And so I'm like, what, how can I get rid of it? So I think I think it would I have like, I do think the one without the notification is more sleek.”

Tree structure:

- At 15:42 min “You could get into rabbit hole with that. Wow.”
- At 15:52 min “I'm sure people who are genuinely interested in the topic would like be like, Oh, wow. And it's very convenient that it's all there.”
- At 16:43 min Participant wants the Title to be more interesting “like catch your titles like this, it is just giving you the summaries and the general information”

Graph Summary

- At 20:12 min get out the key points compare them in one paragraph “some sort of cross paragraph where it's like, this is how they differ this, these are the key points “ “here are the key differences between the two groups and what they think”

Chat feature

- At 22:29 min “maybe not, super often. But if I had like a general question”
- At 22:29 min mentioned that she wants the chat feature as an extension of the tree structure: “Because if you had like, very specific questions, and they weren't in the button options, like the with the tree option, then you might be like, oh, like, I just want to ask my question. But maybe like a combination of the two “
- At 23:28 min the participants wants the LLM summary to act like a person by introducing him the user “That's like, Hi, Lisa is ready to talk to you or something. “
- At 24:19 min she likes the summaries, and it positively helps her to participate in the discussion “I think I like, like the summaries or like, the function of them. I think without them, it would have been like, Oh, what am I looking at”

P5: Male, White, Liberal Arts
Citation:

- At 4:59 min he has more trust in the summary because of the citation “it brings up sort of scholarly articles that show that there's a yeah, there's detailed research into it. It's not just like something that you just came up with on the fly”
- At 5:51 min he would only use citation if it had a hook (sounds interesting) “How, like if the title was provocative in some way, or like, invoked me in some way to interact with it, but it's like, I didn't find it to be interesting that I wouldn't click on it.”
- At 6: 23 min “Yeah. The summary definitely helps. sort of put everything together in one place.”
- At 6:53 min he would change the visual appearance of the summary “with just like this block of text to sign it, it kind of can be a little bit dry and look a little boring.”

Simple Language:

- At 8:46 min he likes the simple language “There's less to process and especially for people who aren't really familiar with the language. Yeah. It'll be a lot easier and straightforward to understand.”

Tree Structure:

- At 14:01 min he feels more in control “You have several options and you can sort of take your own path into what you want to look at.”

Chat feature:

- At 20:55 min he would prefer the tree summary over the chat feature “No, I don't think I would I think I would prefer the model where you could choose the specific topic.”

Participant 7
Citation
He has trust in the source but he would not ready it on a regular basis.
Min 7:55“ Felt overwhelmed by the citation. He thinks it is only for researchers „ like researchers, but if I'm like not doing research, and if I'm just like, like, a regular like customer or like viewer of the page, I think this website is kind of like, yeah, not very intuitive.“
Min 8:41 Would click on it infrequently, especially when he felt that the summary didn’t provide enough information. „I'm getting I will be like, I'll click into ones and like, for now, like, I think I'll just read the summary. i Okay. Yeah. Like, if, if it's giving like enough information, I don't really click into like, any links,“
He likes summary in contrast to no summary
Simple Language Summary
He want the summary to go more in depth on the mood of the conversation
He doesn’t prefer simple language
Participant initiated
When the page loads he wants the. page to signal him that there is a summary button where he could click on.

He likes the notification message red „So yeah, I think I will be like, I think I will be like, um, I will want to click on it. But yeah, like I might like mistake this for like other things. Like,“
The position bottom right corner he thinks more that it might be an FAQ ,Both I think because yeah, like, like basically there's like FAQ something or like frequently asked question like the, like the icon there. Most of them like in the websites are on the bottom right corner. And they all look like like, like the chat icon like this is like a chat and like question mark is like, like, I think yeah, my first thought is like, is it like, an icon that I can? I can type questions in and probably the body will enter me. I think it's more like tastes like from Yeah, based on my intuition so I didn't really realize that it's like the summary like in the beginning if if you don't tell me Okay,‘
Tree Structure
Min 21:32 He likes it more than the other three „this is by far my favorite I'll click onto the link“
Min 21:59 He thinks it is very intuitive / clear „I think it's very intuitive,“
Min 23:25 He thinks it is a very good feature „I think it's a very, like, a good feature. Okay. Okay, yeah.“
Min 23:40 Wants ability to close the summary after he read it and he wants button to jump between the summaries. „example, since like this can be like expanding like infinitely when I clicked into something, I think it's also a good feature to make it bigger like, like, for example, just to cover the web webpage and of course, you have the option to to like close it close the window, but I think you can like make this bigger because if there's like a lot of summaries like after like scroll down or Yeah, or probably there's like some categories or like some like links about like, like I can get into like like just like a link for these but not in this format, but like a more like organized or like structure like format so I can have quick access to like everything.“
Chat feature
Reminds him of Chat bots
Participant 8
Generally
Min 6:26 he would like bullet points „So it would definitely be helpful. Maybe you can like something with the design could be maybe it could be bullet points or something.“
15:32 min he wished the tree structure would be merged with
He wants to know when the AI got he latest data from which year (the same principal as GPT)
Citation
At 7:01 would use citation not very often by like that it is there „not sure if I would use them often, but I wouldn't want to have them there.“
Wants more in depth information
7:20 min likes the styling of the links, intuitively identified the link „I think by now everybody knows that. Like if something is blue and underlined.“
Simple language
8:54 min because he is a foreigner he prefers that simple language „ But I definitely prefer like reading something easier than a more complicated one. “
9:28, is written like you don’t need any prior knowledge it starts with the basics „So it starts with the basics. Where is the audit summary was kind of hard to understand. And I felt like for the other I needed more background knowledge to understand the summary.“
Wants its longer summary, the same style of explanation, but longer.

10:10 min stresses that the simple language is better for foreigners „especially better for foreigners“
He would prefer that summary in contrast to no summary
Participant initiated
At 11:37 min because of the button design and placement he thinks it looks more like a chatbot
He would make the button bigger
He doesn’t think the the current image is representing a summary (he doesn’t have a better suggestion)
Tree summary
14:46 min the tree structure is very intuitive „This is very intuitive.“
14 :54 min because of the button it makes it very intuitive for him, he would interact with the system „I think we all learn to click on buttons and the internet it just looks like every other like random button. So yeah.“
15:18 min he likes the tree structure but he would only use it if he was interested in the topic ,how important the voting is for me, but I think it's very interesting. And you can access a lot of information as fast.“
Graph Summary
He wants a clear contrast between group a and group b
Chat feature
He wants a disclaimer that the information in the chat can be wrong
Participant 9
Likes the chat feature
Participant 10
Generally
He thinks that it looks like an LLM Assistant (he saw it on other websites too) „like pop up boxes, like in websites where they have like, some LM assistant.“
Too positively speaking about UAM, he thinks the LLM is trained with biased data. „I suspect it's because they're probably trained on like, a lot of like little blurbs on websites that are like trying to sell products. So I, I feel like I almost do the skepticism that people have, like, say on a forum more authentically than I would an LM. Like, I feel like this kind of reads to me like it's like a marketing blurb or like something you put in like pitch like venture capital,“
Wants the summary in the system main findings, pro con and state of the art research „So if it could have like, summarize, like the main findings, or like the pros and cons, I think that would be much more like compelling maybe maybe then if like the LLM could like faithfully summarize the like state of the art research on the thing,“
Would give the chat bot the GPT Logo
Simple language
Feels like the uncanny valley „I feel like it's a little bit in the uncanny valley. Almost. When cool use it's like person trying too hard. That's the impression that I get but I do like that. “
Would read this one rather than the citation one
Participant initiated
He wouldn’t click on it because she thinks it is a pop up „associate this with like, sort of like pop ups and like peripheral things that are all around the website.“
Would give the chat bot the GPT Logo

Maybe include a tool tip
Tree Structure
Likes the concept of the tree structure very much
Wants a visual signifier like GitHub „visual signifier that you could use to like show. Okay, this is like the first branch of the tree, it splits off. Here's what's off here. Like right now, I don't think there's any way to didn't note that. “
Graph summary
Wants a visualisation like Obsidian CS Graph with nodes „so there's like a central node that you can make on Yeah, and then you have like ideas branching off of that. I think that would be really interesting to have like an idea of represented by the like, semantic graph like that. You could like click around through it and then like, the LM could like help explain all of the like the different parts of that topic. “
The information provided is overwhelming
Chat feature
Values the chat feature
Participant 11
Generally
Summary to positively about the topic
Said that he is actually missing the links in the simple language
The tree structure is the favourite
Feels like the LLM uses a lot of filler sentence
Citation
He feels like three sources are too much for him
Simple Langauge
It is not deep enough
Feels like the language is too simple
Participant initiated
Likes the participant initiated more
Would place the button in the middle
The red notification message would make her more likely to click on it
Tree structure
Compares it to an „smart FAQ“
Graph summary
Contrast more directly the differences between group A and B
More bullet points
Chat Feature
Likes to get more detailed information quickly
Thinks that the general public can’t ask good questions and therefore the tree structure would be better
Participant 12
Generally
Speaks to positively about it
He wants something that actually doesn’t require an input from the user -> makes it more accessible

Citation
Gets a good general understanding
Simple language
The participant is a tech person and therefore he would wish for more individualisation. He wants a more technical analysis.
Participant initiated
He recogieses the button as an AI chatbot -> he has seen it in other websites too
-> he finds AI chat bots annoying
Tree structure
Wants when a new summary is generated to replace the other prior summary
Thinks because of the button it is very intuitive
Favours that instead of reading long research papers like with the citation feature
Graph summary
He understands, through the summary, pretty good the content of the graph, but he still don't know how to read the graph.
Likes the explanation of the different groups
Chat feature
He likes the ability to directly asks questions
He would like to know which base model llm is used
Participant 13
Generally
He feels like it is biased
Have a sum up sentence in the end
His favourite a the tree and chat feature
Citation
He would rather check on his own for sources. -> „he doesn't trist the technology (LLM)"
He feels like reading an scientific article kills the purpose of a summary
Simple Langauge
Feels like it is more of an introduction than a summary
Feels like it is just written for children
Participant initiated
Compared it to his online banking page, where they have an AI chatbot in the same spot. -> he would expected an AI chat bot
Tree structure
Feels like he get more in depth information now
Loves that he can go deeper into the topic
18:40 min would like sources to the tree struct „I would very much like sources to be their sources again"
Graph Summary
24:12 Is afraid that the LLM hallucinates, because the user doesn't have access to the poll data so even if he wanted to check he can't -> he has to trust the LLM „ So you should you would completely trust this summary. I think it gives you a feeling is little bit uncomfortable because you're trusting a system in some way known for hallucinations and stuff like that to give you summary," -> wants a banner for that

Chat feature
Thinks like it looks like Chat GPT -> therefore he thinks it will have a great acceptance -> claims very intuitively to use
Participant 14
Generally
Make the font bigger for better readability
Citation
He thinks that maybe the citation provided can steer you in a certain way
Simple Langauge
6:59 He says it sounds like a social media post and it would probably resonate better with Gen Z „definitely feels like you're reading it off of social media posts, which definitely would connect to the younger generation for it.“
Participant initiated
Felt like on the page there is a lot of space so the summary should be pulled up initially
Make icon larger
Tree Structure
11:10 gives the user more depth „It's not challenging per se. But it's definitely just more in depth for what they are because you have the tree structure you're able to branch out to and figure out what you want to learn about it“
Have dropdown for the summary
Graph Summary
Camper it with GPT and thinks it is very intuitively

Citation Provided (1 Disagree 5 Agree)

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
I think that I would like to use the summary system frequently.	2.00	5.00	4.33	0.85	0.72	12
I think that I would need the support of a technical person to be able to use this summary system.	1.00	5.00	1.62	1.27	1.62	13
I would imagine that most people would learn to use this summary system very quickly.	4.00	5.00	4.69	0.46	0.21	13
I thought there was too much inconsistency in the summary system.	1.00	3.00	1.77	0.58	0.33	13

Participant Initiated (1 Disagree 5 Agree)

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
I think that I would like to use the summary system frequently.	1.00	5.00	3.77	1.31	1.72	13
I think that I would need the support of a technical person to be able to use this summary system.	1.00	4.00	1.85	1.23	1.51	13
I would imagine that most people would learn to use this summary system very quickly.	4.00	5.00	4.38	0.49	0.24	13
I thought there was too much inconsistency in the summary system.	1.00	3.00	1.67	0.75	0.56	12

Tree Summary (1 Disagree 5 Agree)

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
I think that I would like to use the summary system frequently.	2.00	5.00	4.31	0.82	0.67	13
I think that I would need the support of a technical person to be able to use this summary system.	1.00	4.00	1.31	0.82	0.67	13

I would imagine that most people would learn to use this summary system very quickly.	2.00	5.00	4.23	0.89	0.79	13
I thought there was too much inconsistency in the summary system.	1.00	3.00	1.38	0.62	0.39	13

Graph Summarisation (1 Disagree 5 Agree)

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
I think that I would like to use the summary system frequently.	1.00	5.00	3.77	1.31	1.72	13
I think that I would need the support of a technical person to be able to use this summary system.	1.00	4.00	1.77	1.25	1.56	13
I would imagine that most people would learn to use this summary system very quickly.	2.00	5.00	3.62	1.33	1.78	13
I thought there was too much inconsistency in the summary system.	1.00	4.00	1.85	0.95	0.90	13

Ability to Chat with LLM (1 Disagree 5 Agree)

Field	Min	Max	Mean	Standard Deviation	Variance	Responses
I think that I would like to use the summary system frequently.	4.00	5.00	4.50	0.50	0.25	12
I think that I would need the support of a technical person to be able to use this summary system.	1.00	4.00	1.23	0.80	0.64	13
I would imagine that most people would learn to use this summary system very quickly.	2.00	5.00	4.38	0.84	0.70	13
I thought there was too much inconsistency in the summary system.	1.00	2.00	1.17	0.37	0.14	12

Prompt Summary Evaluation

+++ = very good summary
++ = good summary
+ = moderate summary

Every summary starts with +++.
Deduction exists when:

- The summary doesn't state towards which topics the discussion is positive or negative
- When the summary contains hallucinations or the information provided is wrong
- When the summary doesn't touch all of the basis from the prompt it was give (e.g. it should summarize and say something about group 0 and 1 and the summary doesn't contain anything about the groups feelings)

The summaries were evaluated by Jan Steinhauers opinion.

Group aware consensus + Representativeness of Comments	+++
Summarization of Group A	+++
Summarization of Discussion	+++
Summarization of Discussion	+

Evaluation Refined Summarization Techniques

Summarization Technique A2 Version 1	**++**
Summarization Technique A2 Version 2	**+++**
Summarization Technique A7 Version 2	**+**

Trying and Evaluating prompt strategies from the paper Opportunities and Risks of LLMs f…

Raw Data	++

Polis Summarizing Raw Data

Different summarization techniques	**Summarization of Discussion Topic**	**Summarization of Group A**	**Summarization of Group B**
Outputs	The discussion on Use Cases for Unmanned Aerial Mobility (UAM) highlighted various viewpoints involving the potential applications and acceptance of UAM technology. Common themes that emerged include the interest in utilizing UAM for commuting and package delivery, as well as concerns related to noise levels, safety, and privacy. There is a general agreement on the benefits of UAM for enhancing transportation efficiency and convenience, but divergences exist regarding willingness to pay higher costs, tolerate noise levels, accept frequent aerial presence over residential areas, and overlook safety considerations. Overall, the conversation reflects a mixture of enthusiasm for the possibilities of UAM alongside practical reservations and apprehensions about its implementation.	In reviewing the responses from group A regarding unmanned aerial mobility (UAM), it is evident that there is a level of interest and willingness to explore this innovative mode of transportation. Participants generally agree on the potential for commuting with UAMs, showing openness to the idea of using them for short trips and expressing interest in shifting from traditional ride-sharing services like Uber and Lyft. Moreover, there seems to be a willingness to accept some level of noise and the presence of landing pads in close proximity to personal residences. However, disagreements arise when considering factors such as cost, safety in turbulent conditions, and the constant presence of UAMs directly over personal spaces, indicating concerns about pricing, safety, and potential intrusions into private areas. Overall, group A appears intrigued by the concept of UAM but is	Based on the data provided, it is evident that the participants in group B showcase varying levels of agreement towards the statements related to unmanned aerial mobility (UAM). There is a general consensus within the group in favor of the idea of commuting with UAMs at least once in their lifetime and being open to utilizing UAM services as an alternative to traditional transportation options like Uber, Lyft, or Taxis. However, there is disagreement regarding aspects such as willingness to pay higher prices for UAM rides, tolerance towards potential noise disturbances caused by UAM flights over personal spaces, and feeling safe in UAMs under turbulent conditions. It seems that the group is enthusiastic about exploring the possibilities offered by UAM technology but expresses reservations when it comes to cost, noise impact, and safety concerns.

		cautious about certain aspects, balancing excitement for the technology with practical considerations of comfort, safety, and privacy.	
Jan’s Evaluation	1. The sentence "Overall, the conversation reflects a mixture of enthusiasm for the possibilities of UAM alongside practical reservations and apprehensions about its implementation" captures the overall mood of the discussion, indicating both excitement and concerns about UAM. 2. "Common themes that emerged include the interest in utilizing UAM for commuting and package delivery, as well as concerns related to noise levels, safety, and privacy" shows an understanding of the different perspectives and concerns regarding UAM. 3. "Divergences exist regarding willingness to pay higher costs, tolerate noise levels, accept frequent aerial presence over residential areas, and overlook safety considerations" highlights the	1. "Participants generally agree on the potential for commuting with UAMs, showing openness to the idea" reflects group A's positive attitude towards UAM for commuting. 2. "There seems to be a willingness to accept some level of noise and the presence of landing pads in close proximity to personal residences" highlights their readiness to compromise for UAM benefits. 3. "Disagreements arise when considering factors such as cost, safety in turbulent conditions, and the constant presence of UAMs directly over personal spaces" captures their concerns about UAM. 4. "Group A appears intrigued by the concept of UAM but is cautious about certain aspects" summarizes the group's cautious optimism	1. "Based on the data provided, it is evident that the participants in group B showcase varying levels of agreement...": This sentence sets the context for the analysis, indicating a diverse range of opinions within the group. The clarity here is good, as it prepares the reader for a detailed discussion on these differing views. 2. "...there is a general consensus within the group in favor of the idea of commuting with UAMs...": This statement highlights the positive attitude towards UAM commuting. It is clear and concise, effectively summarizing the group’s overall interest in UAM technology. 3. "However, there is disagreement regarding aspects such as willingness to pay higher prices for UAM rides...": The use of

	range of opinions on the acceptability and practicality of UAM. 4. "There is a general agreement on the benefits of UAM for enhancing transportation efficiency and convenience" summarizes the common ground in the discussion, noting the recognized benefits of UAM. The evaluation shows that the summary effectively captures the dual nature of the discussion on Unmanned Aerial Mobility, highlighting both the optimistic views on its potential and the practical concerns regarding its implementation.	towards UAM technology. The evaluation shows that the summary effectively captures group A's cautious optimism towards unmanned aerial mobility, highlighting their interest, willingness to accept certain downsides, and concerns about cost, safety, and privacy.	"however" introduces the contrasting views well, signaling a shift to the concerns or limitations perceived by the group, which adds depth to the analysis. 4. "...expresses reservations when it comes to cost, noise impact, and safety concerns.": This final point wraps up the group's stance effectively, summarizing the key areas of concern that temper their enthusiasm for UAMs. The evaluation shows that the summary is clear and well-structured, effectively capturing the group's mixed feelings about unmanned aerial mobility.

Different summarization techniques	**Group aware consensus + Representativeness of Comments**	**Summarization Technique A2 Version 2**	**Raw Data**
Outputs	Based on the data provided, participants seem generally open to the concept of utilizing Unmanned Aerial Mobility (UAM) for commuting and express interest in potential cost-effective transportation options, as indicated by the willingness to pay for UAM services at affordable price points. There is a divide in opinions regarding aspects such as UAM delivery services for online shopping packages and the acceptance of UAMs flying over personal spaces, with some expressing agreement and others showing hesitation. Safety concerns come to the forefront when discussing potential noise levels and turbulence during UAM rides, indicating a cautious approach to embracing this innovative mode of transport. Overall, the mood appears cautiously optimistic towards UAM use cases, with a mix of enthusiasm for convenience and apprehension about noise pollution and safety implications.	Detected Topics: 1. Willingness to Use UAM for Commuting and Preference Over Traditional Transport Services 2. Willingness to Pay for UAM Services and Acceptance of Noise Levels 3. Acceptance of UAM Operations in Residential Areas Overall Summary: The sentiment towards unmanned aerial mobility (UAM) varied across the provided comments. Group 0 leaned towards a positive outlook, expressing a high willingness to use UAM for commuting, a preference for UAM over traditional transport services, and a general openness to the idea of UAM operations near residential areas. In contrast, Group 1 displayed more mixed sentiments, with higher levels of disagreement regarding specific aspects such as paying for UAM services, noise	The primary arguments and viewpoints from the discussion revolving around the use cases for Unmanned Aerial Mobility (UAM) suggest a mixed sentiment among users. The common themes or points of agreement among the comments include a willingness to commute with UAMs at least once, the acceptance of UAMs flying over their house or having landing pads nearby, and interest in UAMs delivering online shopping packages. On the other hand, there is general disagreement towards paying higher amounts for UAM services, enduring significant noise levels, or constant UAM flights over personal spaces. Safety concerns, specifically related to turbulence, also emerge as a point of contention. Overall, the discussion highlights a balance between enthusiasm for the potential benefits of UAM and apprehension towards certain practical aspects of its implementation.

		tolerance, and safety concerns during turbulence. Group 1 showed greater hesitation towards fully embracing UAM technology compared to Group 0. Aggregating the agreement percentages, topics 1 and 3 had higher agreement levels in both groups, while topic 2 generated more disagreement in Group 1.	
Jan's Evaluation	1. "Based on the data provided" suggests that the summary is data-driven. This statement establishes that the conclusions are drawn from specific data, indicating a methodical approach to understanding public perception of Unmanned Aerial Mobility (UAM). 2. "Participants seem generally open to the concept of utilizing Unmanned Aerial Mobility (UAM) for commuting" reflects positivity in the participants' attitude towards UAM. This sentence captures the overall positive sentiment towards UAM as a commuting option, showing the initial acceptance of the technology.	1. "Group 0 leaned towards a positive outlook, expressing a high willingness to use UAM for commuting" — This sentence clearly identifies the positive stance of Group 0 towards UAM, making it easy to understand their general attitude. 2. "Group 1 displayed more mixed sentiments, with higher levels of disagreement regarding specific aspects" — This sentence effectively highlights the contrast in opinions between the groups but could benefit from more specific examples to clarify the mixed sentiments.	1. "The primary arguments and viewpoints from the discussion revolving around the use cases for Unmanned Aerial Mobility (UAM) suggest a mixed sentiment among users." This sentence effectively sets the stage for the summary, indicating a balanced discussion on UAM. 2. "The common themes or points of agreement among the comments include a willingness to commute with UAMs at least once..." This sentence clearly identifies the positive attitudes towards UAM, highlighting people's openness to new technology.

	3. "There is a divide in opinions regarding aspects such as UAM delivery services for online shopping packages" highlights the existence of mixed feelings about UAM applications. This shows that while there is interest in the technology, its various use cases are received differently by the public. 4. "Safety concerns come to the forefront when discussing potential noise levels and turbulence during UAM rides" points out specific issues of concern like noise and safety. It indicates that despite the optimism, there are significant apprehensions affecting the acceptance of UAM. The evaluation shows the summary effectively captures the varied responses to UAM, ranging from enthusiasm to concern, reflecting a comprehensive view of public sentiment based on the presented data.	3. "Aggregating the agreement percentages, topics 1 and 3 had higher agreement levels in both groups" — The mention of 'aggregating agreement percentages' suggests a quantitative analysis, but the summary does not provide the actual percentages, which could make the statement hard to interpret. 4. "Topic 2 generated more disagreement in Group 1" — This sentence points out the area of contention clearly but lacks detail on why Group 1 disagreed more on this topic. In summary, the evaluation shows that the summary provides a clear distinction between the groups' perspectives on UAM and highlights key areas of agreement and disagreement. However, it could be more effective with added specific details and percentages to enhance clarity. The data discussed seems comprehensive but sometimes lacks the necessary detail to be fully understandable.	3. "On the other hand, there is general disagreement towards paying higher amounts for UAM services..." This part contrasts the initial enthusiasm by presenting the economic concerns people have about UAM, showing the complexity of public opinion. 4. "Safety concerns, specifically related to turbulence, also emerge as a point of contention." This sentence acknowledges the safety concerns, adding depth to the understanding of public hesitancy. The summary captures the varied public sentiment on UAM effectively but could be more engaging in its presentation of the data.

Comparison of summaries

Group aware consensus + Representativeness of Comments better than Raw Data	The gac + representativeness summary provides a more detailed analysis of the data, including specifics like the willingness to pay for UAM services at affordable price points, showing a nuanced understanding of participant attitudes
	The gac + representativeness summary also overall understands the mood of the participants as "cautiously optimistic," which helps in understanding the general sentiment towards UAM, beyond just the facts of the discussion.
	Both summaries mention mixed sentiments, the gac + representativeness one provides, in my opinion a more balanced view by highlighting both positive interests (cost-effective transportation) and concerns (noise pollution and safety)
Group aware consensus + Representativeness of Comments better than Summarization Technique A2 Version 2	The gac + representativeness summary presents the information in a clear, narrative style that is easy to understand, whereas the second summary is structured more like a report with bullet points, which might be less engaging for a general audience.
	The gac + representativeness summary uses expressive language, such as "cautiously optimistic" and "enthusiasm for convenience," which adds a qualitative depth to the understanding of public attitudes towards UAM.

Summary Results comparison

Summarization of Discussion	The discussion on Use Cases for Unmanned Aerial Mobility (UAM) highlighted various viewpoints involving the potential applications and acceptance of UAM technology. Common themes that emerged include the interest in utilizing UAM for commuting and package delivery, as well as concerns related to noise levels, safety, and privacy. There is a general agreement on the benefits of UAM for enhancing transportation efficiency and convenience, but divergences exist regarding willingness to pay higher costs, tolerate noise levels, accept frequent aerial presence over residential areas, and overlook safety considerations. Overall, the conversation reflects a mixture of enthusiasm for the possibilities of UAM alongside practical reservations and apprehensions about its implementation.
Summarization of Group A	In reviewing the responses from group A regarding unmanned aerial mobility (UAM), it is evident that there is a level of interest and willingness to explore this innovative mode of transportation. Participants generally agree on the potential for commuting with UAMs, showing openness to the idea of using them for short trips and expressing interest in shifting from traditional ride-sharing services like Uber and Lyft. Moreover, there seems to be a willingness to accept some level of noise and the presence of landing pads in close proximity to personal residences. However, disagreements arise when considering factors such as cost, safety in turbulent conditions, and the constant presence of UAMs directly over personal spaces, indicating concerns about pricing, safety, and potential intrusions into private areas. Overall, group A appears intrigued by the concept of UAM but is cautious about certain aspects, balancing excitement for the technology with practical considerations of comfort, safety, and privacy.
Summarization of Group B	Based on the data provided, it is evident that the participants in group B showcase varying levels of agreement towards the statements related to unmanned aerial mobility (UAM). There is a general consensus within the group in favor of the idea of commuting with UAMs at least once in their lifetime and being open to utilizing UAM services as an alternative to traditional transportation options like Uber, Lyft, or Taxis. However, there is disagreement regarding aspects such as willingness to pay higher prices for UAM rides, tolerance towards potential noise

	disturbances caused by UAM flights over personal spaces, and feeling safe in UAMs under turbulent conditions. It seems that the group is enthusiastic about exploring the possibilities offered by UAM technology but expresses reservations when it comes to cost, noise impact, and safety concerns.
Group aware consensus + Representativeness of Comments	Based on the data provided, participants seem generally open to the concept of utilizing Unmanned Aerial Mobility (UAM) for commuting and express interest in potential cost-effective transportation options, as indicated by the willingness to pay for UAM services at affordable price points. There is a divide in opinions regarding aspects such as UAM delivery services for online shopping packages and the acceptance of UAMs flying over personal spaces, with some expressing agreement and others showing hesitation. Safety concerns come to the forefront when discussing potential noise levels and turbulence during UAM rides, indicating a cautious approach to embracing this innovative mode of transport. Overall, the mood appears cautiously optimistic towards UAM use cases, with a mix of enthusiasm for convenience and apprehension about noise pollution and safety implications.
Raw Data	The primary arguments and viewpoints from the discussion revolving around the use cases for Unmanned Aerial Mobility (UAM) suggest a mixed sentiment among users. The common themes or points of agreement among the comments include a willingness to commute with UAMs at least once, the acceptance of UAMs flying over their house or having landing pads nearby, and interest in UAMs delivering online shopping packages. On the other hand, there is general disagreement towards paying higher amounts for UAM services, enduring significant noise levels, or constant UAM flights over personal spaces. Safety concerns, specifically related to turbulence, also emerge as a point of contention. Overall, the discussion highlights a balance between enthusiasm for the potential benefits of UAM and apprehension towards certain practical aspects of its implementation.
Summarization Technique A2 Version 1	In summary, there was general agreement on issues around traffic, public amenities, media/technology, urban planning, and transportation. Views were more mixed on schools, neighborhood issues, healthcare, economics, and community support. A few lighthearted or isolated comments did not show

	clear agreement or disagreement. Overall, the comments indicated interest in improving infrastructure, quality of life, and community support in the area. But there were some polarizing views on select issues like immigrant language ability or reactions to the opioid crisis.
Summarization Technique A2 Version 2	Detected Topics: 4. Willingness to Use UAM for Commuting and Preference Over Traditional Transport Services 5. Willingness to Pay for UAM Services and Acceptance of Noise Levels 6. Acceptance of UAM Operations in Residential Areas Overall Summary: The sentiment towards unmanned aerial mobility (UAM) varied across the provided comments. Group 0 leaned towards a positive outlook, expressing a high willingness to use UAM for commuting, a preference for UAM over traditional transport services, and a general openness to the idea of UAM operations near residential areas. In contrast, Group 1 displayed more mixed sentiments, with higher levels of disagreement regarding specific aspects such as paying for UAM services, noise tolerance, and safety concerns during turbulence. Group 1 showed greater hesitation towards fully embracing UAM technology compared to Group 0. Aggregating the agreement percentages, topics 1 and 3 had higher agreement levels in both groups, while topic 2 generated more disagreement in Group 1.
Summarization Technique A7 Version 2	Detected Topics: 1. Willingness to use UAM for commuting and preference over traditional transportation services. 2. Acceptance of UAM noise levels and willingness to tolerate it in exchange for convenience. 3. Safety concerns related to UAM travel, particularly in turbulent conditions. 4. Views on UAM presence directly over personal living spaces and the acceptance of landing pads nearby. 5. Attitudes towards the cost of UAM travel and willingness to pay varying prices for different routes.

	6. Interest in using UAM for online shopping deliveries. Summary: Topics with high agreement included the willingness to use UAM for commuting and the acceptance of UAM noise levels, indicating general alignment in these areas. Conversely, topics such as safety concerns during turbulence and preferences for online shopping delivery by UAM showed significant disagreement between the two groups. The data suggests that while there is overall enthusiasm for embracing UAM technology for transportation needs and cost considerations, there are differing opinions when it comes to safety and the extent of UAM presence in personal living spaces.

Interview Protocol Version 2

Participant: [_____]
Interviewer: Jan

Hardware:
Mac mini with keyboard and mouse from amazon basic
Macbook Air
Lab monitor, the monitor is showing windows (in chrome) in full screen with 120 % zoom
Iphone as voice recorder

Preparation:
Print out paper that I can record audio Consent to Participate in Research
Mac mini:
Go to Apple website (https://www.apple.com/)
Open 1 Polis topic, without summarization, so the user sees how the default platform looks
Open all 5 Polis topics on different Tabs in the google chrome browser
Open VS Code and check that everything is running properly

Macbook Air (Jan's):
Open Survey on Qualtrics

Part 1: Welcome and Introduction [5 mins]
Hello and welcome! Thank you for participating in our study today. My name is Jan Steinhauer, and I'm conducting research on enhancing deliberation platforms with Large Language Models. Your input will be incredibly valuable in understanding how these tools can be improved to facilitate better discussions and decision-making processes.

Purpose of the Study:
Our study aims to explore how LLMs can be used to summarize discussions on platforms like Polis, especially in the context of Unmanned Aerial Mobility, or UAM. Polis is a digital platform used for gathering and analyzing public opinion. It lets people share their thoughts on various topics and see how others feel about the same issues. We're interested in seeing how these summaries can help users like yourself grasp the essence of discussions more efficiently and make informed contributions.

About the Think-Aloud Method:
We will be using the 'think-aloud' method today. This means we'll ask you to verbalize your thoughts, feelings, and decisions as you navigate through a set of tasks on the Polis platform.

The idea is to speak whatever comes to your mind as you interact with the system. There are no right or wrong answers here; we're interested in your honest experience.

Consent to Participate in Research:
Before we begin, I need to get your consent for the study. The following paper will give you a short overview of the study. Furthermore it explains that the audio will be recorded during the session, as well as your answers in the questionnaire. The recordings will only be used for research purposes to analyze the interaction with the system. Your participation is entirely voluntary, and you may withdraw at any time without penalty. [Jan gives Participant paper "Participate in Research"] You can now read the paper.

[Start recording]
,m
Questions:
Do you have any questions before we get started?

Part 2: *Introduction to Warm-Up Task: [3 min]*
[switch to tab with the apple website]
To get you comfortable with thinking aloud, let's start with a warm-up task. This task is not related to our main study but will help you practice verbalizing your thoughts.

Task Description:
This is the Apple Website. Please go and familiarize yourself with the different Iphone models. As you do this, please verbalize everything you're thinking and doing.

[Participant is browsing through the apple website]
[After 2 minutes there is going to be a hard cut and we move on to the feedback]

Feedback:
Great job with the thinking aloud! You did well explaining your thought process. Remember to keep up this level of detail when we move on to the main task with the Polis platform.

Part 3: Main Task [60 to 80 minutes]

Topic 1
Introduction to Polis:
Now that we've completed the warm-up, we'll begin with the main part of our study. In this section, you'll be using the Polis platform, enhanced with a Large Language Model to help summarize the discussion. Polis is a digital platform used for gathering and analyzing public opinion. It lets people share their thoughts on various topics and see how others feel about the same issues. This tool visually maps out agreements and disagreements within the community, helping to identify common ground and differences. Essentially, Polis is used for engaging

communities in meaningful conversations and helping organizations understand the collective views of their members.

Introduction to Polis Layout:
This is the Polis Website. On the top you can find the title of the discussion. Beneath the title you can find comments, where you can vote Agree, Disagree or Pass [Jan Points to Comments]. If you look at this [Jan points to share your perspective], this is the place where you can generate comments on your own. When you scroll down further you can see a graph, which visualizes what people think about different opinions.

Introduction to UAMs:
The topic of the polis discussions is about Unmanned Aerial Mobility. Unmanned Aerial Mobility (UAM) refers to the development and use of aerial vehicles, like drones and autonomous flying cars, to transport people and cargo without a human pilot onboard.

Setting User Story:
Imagine your city is in consideration to become a test project area for UAMs. Therefore hundreds of citizens and you were selected to comment / vote on UAMs on the polis platform.

Task Description:
[Switch to the first Polis tab]
You'll engage with five different topics on Polis related to UAM. Your task is to mainly interact with the summaries. The summaries are positioned in the bottom right corner [Jan points to the summary]. The summary is about all comments that are in the discussion. Please reflect on their clarity, usefulness, and how the summary may have influenced your understanding of the discussions. Please also express how useful you found the summary, and if you would miss it if it wasn't there. You can start on this page [Jan points to the page that is currently on the monitor].

[Participant interacts with summary and voting]

Follow up questions: Follow up Questions
Thank you for your information. Please fill out the survey on the following pc [Jan points to his laptop. If you have any questions about the survey please feel free to ask me. Furthermore, please tell me when you are finished with it. Moreover please don't do thinking aloud for the questionnaire

[Jan comments out line 43 handleCitationSummaryGeneration() in ConversationUI.js]
[Jan in out line 50 handleSimpleLanguageSummaryGeneration() in ConversationUI.js]
[Jan hits command + s to save the changes]
[Jan switches to third tab with the second topic]

Topic 2 [12 to 17 minutes]
Transition to next Task:
Please move on to the next topic. Remember, we're interested in your continuous thoughts and reactions, especially as you experience the summary.

Follow up questions: Follow up Questions
Thank you for your information. Please fill out the survey on the following pc [Jan points to his laptop. If you have any questions about the survey please feel free to ask me. Furthermore, please tell me when you are finished with it. Moreover please don't do thinking aloud for the questionnaire

[Jan comments out line 50 handleCitationSummaryGeneration() in ConversationUI.js]
[Jan in out line 58 handleSimpleLanguageSummaryGeneration() in ConversationUI.js]
[Jan hits command + s to save the changes]
[Jan switches to fourth tab with the third topic]

Topic 3 [12 to 17 minutes]
Transition to next Task:
That is great. Thank you. Please reflect on their clarity, usefulness, and how the summary may have influenced your understanding of the discussions. Please also express how useful you found the summary, and if you would miss it if it wasn't there.

[Participant interacts with summary]

Follow up questions: Follow up Questions

Transition after the first Task:
Thank you for your information. Please fill out the survey on the following pc [Jan points to his laptop. If you have any questions about the survey please feel free to ask me. Furthermore, please tell me when you are finished with it. Moreover please don't do thinking aloud for the questionnaire.

[Jan comments out line 58 handleCitationSummaryGeneration() in ConversationUI.js]
[comment in either]
[Jan in out line 65 handleGraphsummaryGeneration() in ConversationUI.js]
[Jan in out line 1146 handle_GET_gptSummary() in server.ts]
[or]
[Jan in out line 72 handleDiscussionSummaryGeneration() in ConversationUI.js]
[Jan in out line 1083 handle_GET_gptSummary_tree() in server.ts]
[Jan hits command + s to save the changes]
[Jan switches to fifth tab with the fourth topic]

Topic 4 [12 to 17 minutes]
Transition to next Task:
That is great. Thank you. Please reflect on their clarity, usefulness, and how the summary may have influenced your understanding of the discussions. Please also express how useful you found the summary, and if you would miss it if it wasn't there. The right hand side summaries all the comments

[Participant interacts with summary]

Follow up questions: Follow up Questions
Thank you for your information. Please fill out the survey on the following pc [Jan points to his laptop. If you have any questions about the survey please feel free to ask me. Furthermore, please tell me when you are finished with it. Moreover please don't do thinking aloud for the questionnaire

[Jan comments out line 58 handleCitationSummaryGeneration() in ConversationUI.js]
[Jan in out line 195 rcw-sender in ConversationUI.js]
[Jan switches to sixth tab with the fifth topic]

Participant finishes survey:
Hey I am ready with the survey.

Topic 5 [12 to 17 minutes]
Transition to next Task:
That is great. Thank you. Please reflect on their clarity, usefulness, and how the summary may have influenced your understanding of the discussions. Please also express how useful you found the summary, and if you would miss it if it wasn't there. The right hand side summarizes the discussion
[Participant interacts with summary]

Follow up questions: Follow up Questions

Thank you for your information. Please fill out the survey on the following pc [Jan points to his laptop. If you have any questions about the survey please feel free to ask me. Furthermore, please tell me when you are finished with it. Moreover please don't do thinking aloud for the questionnaire

Topic 6 [12 to 17 minutes]
Transition to next Task:
That is great. Thank you. Please reflect on their clarity, usefulness, and how the summary may have influenced your understanding of the discussions. Please also express how useful you

found the summary, and if you would miss it if it wasn't there. The right hand side summarizes the discussion
[Participant interacts with summary]

Follow up questions: Follow up Questions
Thank you for your information. Please fill out the survey on the following pc [Jan points to his laptop. If you have any questions about the survey please feel free to ask me. Furthermore, please tell me when you are finished with it. Moreover please don't do thinking aloud for the questionnaire

End of Experiment [2 minutes]
Conclusion of Main Tasks:
Thank you for completing the main tasks. Your feedback is utterly valuable to our research. Thank you for completing the survey. Before we conclude, I'd like to take a moment to debrief you about the study and ensure any questions you have are answered.

Explaining the Purpose Again:
As mentioned at the beginning, the purpose of our study is to explore how Large Language Models can be integrated into deliberation tools like Polis to enhance the deliberation process through effective summarization of discussions, specifically in the context of Unmanned Aerial Mobility.

Highlighting the Importance of Their Participation:
Your participation today has been incredibly valuable. The insights you've provided will help us understand the potential of LLMs in making community deliberations more accessible and engaging for everyone involved.

Answering Participant Questions:
Now, I'd be happy to answer any questions you might have about the study, our methods, or the potential implications of our findings. Please feel free to ask anything.

Add further prompting or backup questions instructions if you find that the interview ends early/participant doesn't explore all parts of the website, etc

1. Do you understand the concept of UAMs?
2. Please tell me what the last discussion was all about.
3. Were there terms, which you didn't understand?
4. Is there a reason why you didn't engage with the summary?
5. What has changed between the summary of the last topic in comparison to this topic? -> Which way did you like it better?
6. Is there something that bothered you about the summary?
7. Is there something you would change about the summary?

I personally think that if a participant doesn't explore all parts of the website then maybe it just shows a UX issue. I would ask the participant: Hey [name], have you seen the [fe ature that he didn't use]? If the answer is yes then I would ask him why he didn't interact with it. If the answer is no I would note it down as an issue.

Interview Questions about how good the summary is:

1. Having read the summary, do you think you have gained a vivid understanding of how people feel about [the topic]?
2. Taking into consideration the graph, do you think you now have a better understanding of how to interpret the graph, with the summary?
3. Taking into consideration the group A and B, do you think you now have a better understanding of how to interpret both groups, with the summary?
4. Did seeing the summary change your initial opinion? -> then question 5 e summary? (peer pressure) did you feel like the summary better prepared you?
5. Did the summary provide a balanced view of [the topic] as discussed in the poll, or did it seem biased toward one perspective in poll data?
6. After reading the summary, do you feel more confident in discussing [the topic] with others?
7. Did the summary's tone or language use affect your perception of [the topic]?
8. Did the summary successfully engage your interest and maintain your attention throughout the reading?
9. Were there any areas in the summary where you felt additional information or context would have been beneficial?

Why Thinking Aloud

1. Cognitive Transparency: The think-aloud protocol provides insights into the participant's thought processes, making it easier to understand how they interpret and make decisions based on the summaries provided by LLMs. This method is especially valuable for revealing gaps in information, misunderstandings, and the decision-making process, which are not easily captured through other data collection methods such as surveys or observational studies (Ericsson & Simon, 1993).
2. Usability and Interaction Issues Identification: It helps in identifying usability problems within the digital tool. Participants vocalize their thoughts, feelings, and confusion as they navigate through the Polis platform, interact with the LLM-generated summaries, and participate in discussions. This real-time feedback is crucial for identifying design and interaction issues that may not be apparent to designers or through quantitative data alone (Nielsen, 1993).

Ericsson, K. A., & Simon, H. A. (1993). Protocol Analysis: Verbal Reports as Data. MIT Press.
Nielsen, J. (1993). Usability Engineering. Academic Press.