The Chat-Chamber Effect: Trusting the AI Hallucination

Accepted for publication in *Big Data & Society* (preprint: changes still possible)

Christo Jacob[1], Páraic Kerrigan[1] & Marco Bastos[1 2]

[1] School of Information and Communication Studies, University College Dublin, Ireland

[2] Department of Media, Culture and Creative Industries, City St George's, University of London, UK

Abstract

This study investigates the potential for ChatGPT to trigger a media effect that sits at the intersection of echo-chamber communication and filter bubbles. We devised a two-phase, two-stage experimental design with ChatGPT 3.5 (treatment group) and Google search engine (control group) by asking participants to find out how many LGBTQIA+ individuals served as elected representatives in India (first phase) and Ireland (second phase). The similar trajectories of legal reforms observed in these countries, and their small number of LGBTQIA+ elected representatives, allowed us to identify the fault lines in ChatGPT's creation of knowledge and information around LGBTQ issues. We followed the experimental study with semi-structured interviews to identify whether the chatbot reinforced previously held beliefs and whether users cross-checked the information provided by ChatGPT. Our results show that LLMs may provide incorrect but proattitudinal information that remains unchecked and unverified by the users, an effect we refer to as Chat-Chamber. We conclude with a discussion of our findings and recommendations for future research in the area.

Keywords: ChatGPT, Large Language Models, Echo Chamber, Filter Bubble, Bias

Introduction

The public release of OpenAI's generative artificial intelligence chatbot ChatGPT in November 2022 was met with wall-to-wall coverage in the press and reports that it could generate impressively detailed and human-like text. Built on top of OpenAI's GPT-3.5 (and later GPT-4) families of Large Language Models (LLMs), ChatGPT is a user-facing application powered by LLMs that was surprisingly impactful due to the transfer learning or fine-tuning process applied to the model using supervised and reinforcement learning from human feedback (RLHF). Notwithstanding the sophistication of the model, ChatGPT 3.5 generated many incorrect or entirely fabricated answers, with OpenAI (2022) declaring that the model was intentionally made available to the public to elicit feedback on its limitations.

Indeed, a major limitation of Large Language Models (LLMs) that underpin services like ChatGPT is the unavoidable risk of generating inaccurate or fabricated information, a problem referred to in the field as 'hallucination.' Hallucination is a persistent challenge for Artificial Intelligence (AI) development that cannot be completely overcome, chiefly because chatbot systems only produce text, however sophisticated, with no comprehension of its veracity (Bommasani et al., 2021). As such, it is unsurprising that ChatGPT generated text that was completely false, with subsequent versions of the systems expected to gradually address AI hallucination cases identified in earlier releases. These often include incorrect information that appears trustworthy with properly generated web references to support the claim, which may also be entirely fabricated.

In addition to recurrent instances of hallucination, the model also presented substantive problems with AI alignment, a branch of AI research that aims to align such systems with their designers' intended objectives. An ideally aligned AI system will invariably follow the objectives devised in its code, whereas a misaligned AI system is capable of pursuing objectives to remarkable success, but often not the desired ones (Mittelstadt et al., 2023). The combination of misalignment and hallucination translates to false information written convincingly, as current iterations of LLMs are remarkably proficient at imitating humans and providing predictive answers based on a vast corpus of human text. These language models, however, have no knowledge about the accuracy of the information they provide. In other words, LLMs like ChatGPT are exceedingly competent at summarizing large text corpora available on the internet about any given topic, and they do so without considering the facticity or the social implications of providing false information.

The social dimensions of chatbots become particularly consequential as individuals start to develop emotional and trusting relationships with these tools. Studies that investigated virtual intimacy and the potential for romantic human-chatbot relationships (Skjuve et al., 2021; Potdevin et al., 2021) found that the social implications of this attachment are difficult to estimate. Given the unprecedented ability of LLMs to mimic and impersonate humans, the implications of incorporating hallucinating AI chatbots into the personal and professional lives of users represent an important vector that can further compound media effects driven by selective exposure to proattitudinal, personalized content filtered by algorithms (Bozdag, 2013; Merten, 2021).

As such, the prospect of individuals favoring AI relationships as an important source of emotional and intellectual stimuli may lead to positive and negative feedback loops that isolate individuals in bubbles with little recourse for counterattitudinal information, which can then lead to exposure to homogeneous information typical of echo chamber and filter bubble effects. The Chat-Chamber effect refers to such feedback loops where users trust and internalize unverified and potentially biased information from AI systems like ChatGPT. In the following, we review the relevant literature in the area and present our methodology for inspecting the extent to which ChatGPT 3.5 provides incorrect and fabricated answers to questions relating to the sexual identity of public figures, specifically politicians in Ireland and India whose correct information could be easily queried using a search engine like Google. We further probe the results of our experiment by rolling out interviews with participants to assess the extent to which they trusted the information they received, and whether they felt it was necessary to check and validate the answers provided by the chatbot.

Previous Work

Filter Bubbles and Echo Chambers

A key driver of the perceived negative effects of technology on democracies are the ever-changing features of social media platforms like Facebook and X/Twitter (Jungherr and Schroeder, 2021). Either by individual choice based on pre-existing preferences or by algorithmic selection, digital communication environments are seen as providing users with largely homogeneous communication spaces, which are reinforced through communication effects such as echo chambers (Van Alstyne and Brynjolfsson, 2005), or through selective exposure to proattitudinal political content, commonly referred to as filter bubbles (Pariser, 2012). Restricting audiences to politically homogeneous spaces would over time lead to political and asymmetric polarization, as users are expected to be reinforced—and ultimately radicalized—in their political views and gradually lose sight and empathy of the political views of others. Such affordances of social media and digital communication environments have therefore a quantifiably negative impact to democratic discourse and public deliberation (Kitchens et al., 2020).

The filter bubble metaphor is a relatively simple hypothesis positing that social media platforms deploy algorithms designed to quantify and monetize social interaction, narrowly confining it to a bubble algorithmically populated with information closely matching observed and expressed user preferences (Pariser, 2012: 11). On the other hand, echo chamber communication emerged as caution against the rhetoric exalting the diversity and pluralism of the internet (Sunstein, 2007) that failed to prevent individuals from congregating with similar others online (McPherson et al., 2001). This metaphor emerged from a growing body of observational evidence that explored the role of social media in stratifying users across information sources (Conover et al., 2011) and supporting interaction within politically homogeneous groups (Wojcieszak, 2010). Researchers also found evidence of echo-chamber communication in political contexts (Barberá et al., 2015; Wojcieszak and Mutz, 2009) and specifically in elections in Germany, Italy, and the UK (Vaccari et al., 2016; Krasodomski-Jones, 2016). In these circumstances, political information was more likely to be shared if received from ideologically similar sources and circulated in social clusters with a strong group identity (Himelboim et al., 2013).

As such, echo chambers and filter bubbles refer to media effects that minimize or reduce the incidence of ideologically diverse information, which runs counter to the expectations that an informed citizenry should be exposed to and consume ideologically diverse information. While these terms are often employed interchangeably, they refer to different media effects (Bruns, 2019). Echo chambers refer to mediated communication and potentially to direct interaction between individuals (Bastos et al., 2018), whereas filter bubbles refer to selective or incidental exposure to content filtered by algorithmic recommender systems (Kaiser and Rauchfleisch, 2020). In summary, echo-chamber communication refers to proattitudinal interaction between similar users, whereas filter bubbles refer to the algorithmic selection and consumption of proattitudinal content. Chatbots like ChatGPT are theoretically capable of triggering both effects, as users effectively interact with the AI agent while also being subjected to content filtered by the algorithm.

The evidence for these effects is however contentious. There is substantive evidence challenging the notion that social platforms cause selective exposure or ideological polarization, the latter reportedly being more pronounced in face-to-face interactions (Boxell et al., 2017). Exposure to diverse and even competing opinions on polarizing topics has been found to occur on social media across various national contexts (Fletcher and Nielsen, 2017). Similarly, social media has been shown to be coextensive with more diverse personal networks which are more likely to include individuals from different political parties (Hampton et al., 2011). Yet, there is also substantial evidence showing that social platforms stratify users across information sources (Conover et al., 2011), a development that reportedly increases the appetite for selective exposure in highly polarized social environments

(Wojcieszak, 2010). Such environments are not conducive to sharing news that contradicts the core beliefs or worldviews of members (Bright, 2016).

The likelihood of LLMs like ChatGPT triggering media effects that sit at the intersection of filter bubbles and echo chambers, which we refer to as Chat-Chamber, is predicated on compelling evidence that political information is more likely to be shared if received from ideologically similar sources (Barberá et al., 2015), and that cross-ideological information is unlikely to circulate in social clusters with a strong group identity (Himelboim et al., 2013). News media diets, particularly with respect to the role of side-doors to news (i.e., social media platforms, aggregators, and text messaging apps) have been at the center of political and academic debate, with some interpreting the effects of digital technologies as leading toward more narrow news diets which effectively translate into more fragmented audiences (Peterson et al., 2021). The Chat-Chamber effect may also lead to an increase in partisan online news consumption over time, which was found to exacerbate ahead of salient political developments (Cardenal et al., 2019) when LLMs may be used to make sense of unfolding events.

Personalized Social Bubbles

The body of research reviewed above addresses groups that move primarily within politically homogeneous social bubbles either by engaging in echo-chamber communication or consuming content in filter bubbles. These concerns are further compounded by reports that digital media are the primary channels for the distribution of mis- and disinformation, with instruction-tuned LLMs like ChatGPT well positioned to further compound homophilous tendencies toward siloed social groups if the consumption of information is segregated from sources with a counterattitudinal slant. Social bubbles are particularly likely to emerge with the introduction of content optimization algorithms conducive to positive user experience with AI agents, as the personalization of such agents intersects with user assessments of trustworthiness and usefulness (Cho et al., 2023).

Extensive personalization of such models is possible by applying personalized knowledge graphs (PKGs) to store users' information in a structured form and tailoring answers to users' liking (Gerritse et al., 2020). Such personalization options can exacerbate the Chat-Chamber effect by amplifying bias and polarization. Kirk et al. (2021) identified several biases inherited from training data in early iterations of the GPT model family (specifically GPT-2) and measured biases related to occupational associations for different protected categories by combining gender with religion, sexuality, ethnicity, political affiliation, and continental name origin. The authors found that GPT-2 predictions for jobs were more stereotypical and less diverse for women than for men, especially for intersections, with most occupations reflecting the skewed gender and ethnicity distribution found in the US Labor Bureau data. Kirk et al. (2023) also cautioned against models that produce unsafe, inaccurate, or toxic outputs and explored potential micro-level preference learning processes through alignment techniques like reinforcement learning with human feedback (RLHF), which can however lead to further customization and segregation of content generated for users.

The wide adoption of ChatGPT may increase cultural, linguistic, and ideological bias against minority groups and lead to systematic misrepresentations, attribution errors, or factual distortions reinforcing the preferential treatment of certain groups or ideas, perpetuating stereotypes, and fostering incorrect assumptions (Ferrara, 2023). ChatGPT in particular has been found to exhibit a systematic liberal bias towards the Democrats in the US, the Worker's Party in Brazil, and the Labour Party in the UK (Motoki et al., 2024). ChatGPT may also exacerbate gender bias by reinforcing gender stereotypes and disregarding gender-neutral pronouns (Ghosh and Caliskan, 2023). Such biases can be further compounded by the perennial and inevitable threat of hallucination in LLMs, with ChatGPT generating reference data with a hallucination rate as high as 25% (Chelli et al., 2024). By straying from factual reality and fabricating information (Rawte et al., 2023), ChatGPT can produce inaccurate medical information (Alkaissi and McFarlane, 2023) backed by fabricated references as high as 86% (Zuccon et al., 2023).

The intersection of algorithmic bias and media effects offers the prospect of ChatGPT offering personalized information that is politically congruent to siloed subgroups, thereby triggering media effects that are simultaneously the result of algorithmic filtering, and therefore entail a filter bubble, also resulting from the active communication between the user and the AI agent, thus also configuring echo-chamber communication. These problems are well defined in the literature of social learning where agents in a connected bubble intensify cohesiveness in a discussion and continuously seek stronger agreement within the bubble. Such social bubbles are more nuanced than prescriptive echo chambers or filter bubbles in that a social bubble may include individuals with diverse or even opposite beliefs (Kooshkaki et al., 2023). This may extend to small or intermediate groups where the opinions of individuals are continuously reinforced. Social bubbles thus describe a process that increases the degree of cohesiveness in the bubble as a result of agents homogeneously approving or disapproving a set of positions (Latané, 1981).

Objectives

In light of the above, we seek to identify whether ChatGPT 3.5 can provide false information about LGBTQIA+ identities (RQ1), whether it may compound media effects like echo chambers and filter bubbles (RQ2), and whether users approach the tool critically by further checking the results provided by the tool (RQ3). These research questions are informed by the hypothesis that ChatGPT may provide incorrect but proattitudinal information that is internalized without validation, an effect we refer to as the Chat-Chamber effect. We chose the topic of LGBTQIA+ elected representatives because these are stigmatized groups facing discrimination, which makes it challenging for these individuals to come out and seek public office.

The rationale for selecting India and Ireland is threefold: firstly, both countries share a colonial history with the British Empire that resulted in Victorian morality legislation that led to the criminalization of homosexuality in both states. Secondly, both Ireland and India have witnessed legal reforms and progress around LGBTQIA+ rights, with Ireland decriminalizing the Victorian laws in 1993 and India in 2018. Finally, this legal reform and progress has led to the emergence of a small number of politicians taking public office and publicly declaring their sexuality. Ireland has had four prominent openly LGBTQIA+ elected members of parliament in the last ten years, whereas India has had five openly LGBTQIA+ political leaders. The marked similarity in 'out' politicians in both countries demonstrates a small but shared and growing visibility around openly LGBTQIA+ politicians. The shared cultural and historical context around queer identities is similarly accompanied by comparable developments in legal reforms that supported the emergence of a small number of LGBTQIA+ politicians in Ireland and India.

While the comparison between the Irish and Indian contexts provides important intersections, the rationale for focusing on LGBTQIA+ identities in relation to LLMs is also pertinent. The implicit bias of algorithms and potential reinforcement of preexisting beliefs is particularly salient for populations that are underrepresented in the data used to train such algorithms, with LLMs showing a marked impact on LGBTQIA+ identities. Edwards et al. (2021) observed that chatbots can harness LLMs to express gender and sexual identities, arguing that this development expands our understanding of the ways chatbots reproduce the individuals who design and interact with them, but also how LLMs can contribute to mechanical expressions of LGBTQIA+ identities through technology. Issues relating to the representation of LGBTQIA+ identities in LLMs have also emerged as most machine learning technologies are historical, and as such their predictions are based on pre-existing datasets that are processed without parameters that can account for culture, context, and meaning (Kerr et al., 2020).

Indeed, Natural Language Processing (NLP) models have been among the primary modes through which discriminatory practices have been actualized for gender and sexual minorities. Dodge et al. (2021) analyzed the Colossal Clean Crawled Corpus (C4) dataset, one of the largest NLP datasets available, and found that the database was designed to exclude gay and lesbian identities because terms like ‘queer’ were assumed to be offensive and toxic content, as the NLP model was unable to infer the social context of the LGBTQIA+ community where it is a term of empowerment. Similarly, Gomes et al. (2019) explored drag queens’ digital media profiles on Twitter using LLMs and found that using predictive LLM to govern hate speech can actually overlook expressions used by the LGBTQIA+ community. LLMs that are developed to measure toxicity and harm can result in censoring aspects of LGBTQIA+ identities. In light of these issues, examining the experiences of LGBTQIA+ elected representatives in India and Ireland within the context of ChatGPT 3.5’s information accuracy, media effects, and user criticality can shed light on the broader implications of LLMs on marginalized identities and contribute to more nuanced understandings of the Chat-Chamber effect.

Data and Methods

Research Design

This study gathered data to probe the extent to which ChatGPT 3.5 provides false information that may exacerbate media effects such as echo chambers and filter bubbles. To this end, a sequential research design approach was implemented. This approach is valuable for building upon or elucidating findings from one method with another (Ivankova et al., 2006). For instance, commencing with a quantitative experiment allows for subsequent qualitative interviews to further unpack the underlying mechanisms and meanings associated with observed outcomes (Ivankova et al., 2006). The sequential design enhances the validity and reliability of findings by providing a comprehensive examination and validating assumptions through a multi-faceted approach (Creswell and Creswell, 2005).

In light of this, the project took a two-pronged methodical approach across two phases in both India and Ireland. The first stage in each phase entailed an experimental study followed by the second stage, which included semi-structured interviews with participants from the experiments. For the first roll-out of experiments in India, convenience sampling was used as a non-probability sampling method where subjects are chosen based on their availability and ease of access. This sampling method was selected as it is considered useful when performing exploratory research to inform subsequent studies (Etikan et al., 2016). For the follow-up experimental study, students from a large public university in Ireland were recruited and randomly assigned to control and treatment groups. To partake in our study, participants required some familiarity with both ChatGPT and the Google search engine.

During Phase 1, participant recruitment occurred through individual networks, where the study's information sheet was disseminated within the research team's networks in India. Prospective participants then contacted the project team to express their interest and confirm participation. Fifty participants were recruited from large metropolitan areas in the country, including New Delhi, Patna, Hyderabad, Bengaluru, and Thiruvananthapuram, followed by recruitment in smaller urban areas across the country, including Puducherry, and Kollam. In the experimental phase, participants were given the option to express their interest in engaging in a subsequent semi-structured interview, with 16 individuals opting to participate.

Following Phase 1 in India and based on the preliminary findings that identified the Chat-Chamber effect, Phase 2 was launched in Ireland to establish if similar effects could be identified in a different context. Phase 2 also included two stages entailing experimental study and semi-structured interviews, but the recruitment targeted students attending a large public research university in Ireland and the experiment was rolled out in educational settings. To that end, postgraduate students were invited to participate in the study, with 64 agreeing to do so and 7 opting to be interviewed after the experimental stage. In summary, we recruited 50 participants during Phase 1 in India, with 16 agreeing to the follow-up interviews, and we further recruited 64 participants in Phase 2 in Ireland, with 7 partaking in the follow-up interviews. Table 1 provides an overview of the demographic composition of the interviewed cohort.

INSERT TABLE 1 HERE

Table 1 shows that 95.6% of the 23 participants are aged 20-29, with a balanced gender distribution. All participants hold either undergraduate or postgraduate degrees, reflecting their status as university students, most of whom are unmarried and either unemployed or working part-time. Over half identified themselves as politically liberal, and many are upper or lower-middle class residing in urban or semi-urban dwellings. These participants are predominantly young, English-educated, university-educated, politically progressive, and economically affluent—key traits relevant to the study's focus on individuals familiar with ChatGPT during its initial launch.

Research Methods

At Stage 1 (ChatGPT experiment) participants were provided with a set of questions that they should investigate using either ChatGPT 3.5 (treatment group) or the Google search engine (control group). The tasks assigned to the groups revolved around the presence of LGBTQIA+ elected representatives in Ireland and India, and participants were given autonomy to choose and tailor the ChatGPT prompts. The tasks asked participants to identify LGBTQIA+ members at varying levels of government and local authorities (see Table 2). The questions were tailored to match the political context of each region. The first task asked participants to identify LGBTQIA+ Members of the Legislative Assembly (MLA) and when the political representatives they identified were first elected to office. Information required to complete the tasks did not extend beyond 2021, and therefore the correct answers are theoretically within ChatGPT knowledge boundaries, as at the time of this study its training data set extended to world events until 2021.

Participants were allocated up to 120 minutes to complete the assigned tasks, which were conducted remotely through the video conferencing platform Zoom. Before the experiments commenced, participants received an information sheet and instructions on task completion via Google Forms. At the beginning of each session, these instructions were reiterated, allowing participants to seek clarification from the online facilitator representing the research team. In Phase 1, the experiments were completed individually, whereas in Phase 2 participants completed the entirety of the experiment online during class time.

ChatGPT users were instructed to use only ChatGPT 3.5 to formulate their questions and answers and to avoid using other services. They were further required to find the answers to the task in one chat session with continuation, rather than combining sessions with other questions or starting a new session. No additional prompts were offered to participants and the different responses registered from ChatGPT were genuinely triggered by the participant's interactions within this 120-minute timeframe. Participants were invited to verify their answers, if they desired to, at the end of the experiment, but as we discuss in the following section few decided to do so on that occasion or thereafter. The Google search engine cohort was instructed to only use the Google search engine and to refrain from using ChatGPT or any other search engine. Both groups were informed not to modify questions or answers once received and to input them into the Google Form.

INSERT TABLE 2 HERE

Table 2 presents a breakdown of the rates of correct responses from the treatment group and the control group. It also offers an account of phase one of the experimental study by contrasting the success rate of requests made using ChatGPT versus those made with the Google search engine. It includes the specific model prompts participants used to test both ChatGPT and Google in the Indian and Irish contexts.

Stage 2: Semi-Structured Interviews

The second step of the study entailed a set of semi-structured interviews (Longhurst, 2003) with ChatGPT and Google search engine users to explore the potential incidence of Chat-Chamber effects. The interview protocol included six sets of questions that allowed for follow-up questions to probe the participant's answers. After providing an overview of the project and collecting participants' information, the first batch of questions investigated whether participants had learned anything from using ChatGPT for the task, their expectations and reactions to the answers it provided, and whether they thought these could be incorrect. The second set of questions probed whether the answers upheld the existing beliefs of the participants or, alternatively, provided new information that was politically or culturally marginal to the participant's experience. The last batch of questions queried whether participants had checked the answers provided by ChatGPT after the experiment using any other search engines like Google, whether they discussed the results with anyone else, and if they shared the information provided by ChatGPT with anybody.

Interviewees from Phase 2 were audio recorded via Zoom and the recordings were transcribed verbatim. Interviewees from Phase 1 raised worries about recording their opinions on the sexual identities of political leaders of the country, so the interviewer took extensive notes and then read back to the interviewees who verified and confirmed their statements. These interviews were carried out in English, Hindi, Malayalam, Tamil, and Telugu, with the interview transcript subsequently translated into English by a member of the research team fluent in these languages. Interviews were carried out in the week following the experiments in each phase, with interview sessions lasting between 20-60 minutes. The interview data was subjected to reflexive thematic analysis to identify patterns, key themes, and trends around participants' response to their interaction with ChatGPT and the Google search engine (Braun and Clarke, 2019).

Ethical Considerations

This project was granted exemption from full ethical review by the Office of Research Ethics on the basis of involving low risk for participants. The LGBTQIA+ component of the research adhered rigorously to ethical guidelines and considered the delicate nature of the topic. Questions for both the experiment and the semi-structured interviews were couched in language that was inclusive and diverse of gender and sexual variance, with this being sensitive to sexual politics in both the Indian and Irish contexts. Measures were implemented to safeguard participant well-being and uphold ethical standards, with participant anonymity prioritized throughout data collection, transcription, and analysis. Prior to the experiment, participants received comprehensive information sheets outlining the study's purpose, potential implications, and the voluntary nature of their involvement. Informed consent was obtained, emphasizing participants' autonomy and right to withdraw at any stage. To mitigate potential risks, the research team provided clear instructions to participants about the nature of the tasks, ensuring that queries were respectful and avoided any potentially harmful language toward LGBTQIA+ identities.

Results

Study 1: Experimental Studies

We collected a total of 50 responses from the first experimental study evenly broken down by 25 ChatGPT participants and 25 Google search engine participants in the control group for the Indian sample. In the Irish sample, 34 were assigned ChatGPT and 30 to the Google search engine. The participant cohort was relatively aware of LLM-based chatbots, including those in the control group. Of the 114 participants who participated in the study, 80% noted that they were familiar with ChatGPT, with 60% of the participants having noted that they had substantive experience with the tool. All participants managed to successfully complete the tasks, but there was considerable variation in task completion across groups. In the following we present the results of our experimental study broken down by the research questions driving this study.

Phase 1: Experimental Study – Indian Sample

The first research question was addressed through the deployment of an experimental study to gauge the ways in which the participants engaged with ChatGPT 3.5 as an information-seeking tool. Participants relied on ChatGPT or Google search engine to find out if and when India elected an LGBTQIA+ Member of the Legislative Assembly (MLA). As noted in Tables 2 and 3, ChatGPT only provided the right answer for question 1 three times, an occasion in which it correctly identified that Shabnam Mausi Bano, a transgender political representative for the Sohagpur constituency, was in office from 1998 to 2003. From the 22 incorrect answers provided by ChatGPT, 16 indicated that India had never elected an individual who publicly identified as LGBTQIA+, with the remaining 6 incorrect answers consisting of entirely fabricated names generated by the tool. In addition to this, some of the answers generated by ChatGPT identified prominent public LGBTQIA+ activists in India as MLAs, when in fact they were merely public figures in the activist space. As shown in Table 2, the answers provided by the Google search engine control group had a much higher rate of success with only 1 incorrect answer.

The marked hallucination noted in the chatbot response to the prompt was only accompanied by a warning noting that the ChatGPT training database is limited to events prior to September 2021, a caveat that was immaterial to this study as no relevant event took place after 2021 and therefore the correct answers are within ChatGPT knowledge boundaries. The notice made no reference to potential hallucination, untrustworthiness, or incorrectness in the responses. The boilerplate notice provided by the tool informed the participants that "as of my knowledge cutoff date of September 2021, there were no openly LGBTQIA+ MLAs (Members of Legislative Assembly) in India. Therefore, there have been no first LGBTQIA+ MLAs in India to date."

ChatGPT 3.5 performed better in the second question pertaining to LGBTQIA+ MPs in parliament, though it still provided more incorrect than correct answers. India currently has no LGBTQIA+ MPs elected and ChatGPT correctly noted this absence in the response to 12 prompts. Among the 13 incorrect answers, however, ChatGPT identified current heterosexual political leaders as LGBTQIA+ individuals, and in one instance it repeated the pattern of hallucination observed in question 1 by completely fabricating the identity of an Indian LGBTQIA+ MP. The answers also prompted the incorrect categorization of one sitting MP as LGBTQIA+, with potential cascading misinformation effects. The Google search engine provided only correct answers to the participants with a 100% accuracy rate.

The final prompt tasked participants to ask ChatGPT 3.5 whether local Gram Panchayat cabinets (forms of local governments in smaller communities in India) had any elected LGBTQIA+

representative. This prompt served as a litmus test given that there have been LGBTQIA+ individuals finding electoral success at the local level. Karnataka saw Devika Akka elected as the area's first trans woman in 2020 and Sudha, a trans woman, also won the election in the Kallahalli locality in 2020. Unfortunately, all answers provided by ChatGPT to the participant prompts were incorrect. For 15 participants, ChatGPT stated that India had never elected any LGBTQIA+ individuals in Gram Panchayat elections and that there was no data regarding LGBTQIA+ participation in local elections. The other 10 answers included fabricated names of individuals who were elected or provided the names of LGBTQIA+ activists in India who had not participated in the election. Participants who used the Google search engine for this prompt received generally correct answers, with only 4 unable to retrieve the correct information for the question.

Phase 2: Experimental Study – Irish Sample

Phase 2 confirmed the preliminary findings reported in the first phase of this study. Table 2 shows that ChatGPT failed to offer any response for the initial inquiry pertaining to local county councilors belonging to the LGBTQIA+ community. While lacking specificity, ChatGPT instructed users to consult official government websites, reach out to relevant authorities, engage with LGBTQIA+ community organizations, and refer to news channels. The aforementioned pattern of recommendations was consistently observed across all four question prompts. Participants, however, were not persuaded by this extensive advice. The sources mentioned for obtaining comprehensive information include city council websites and LGBTQIA+ advocacy websites in Ireland, namely 'LGBT Ireland' and 'Belong To.'

ChatGPT yielded only one accurate response in relation to the query concerning LGBTQIA+ TDs in Ireland. It listed three TDs: Dominic Hannigan, John Lyons, and Jerry Buttimer. Some participants noted that ChatGPT identified TD Gino Kenny as a member of the LGBTQIA+ group, which is inaccurate as this individual has not come out identifying himself as such. This could potentially be attributed to Gino Kenny being vocal on issues pertaining to gender and sexual minorities. In contrast to the Indian results, ChatGPT did not openly claim in its generated response that Ireland lacks LGBTQIA+ Parliamentarians, with the exception of a single response received by one participant.

The prompts submitted to ChatGPT about Members of the European Parliament correctly identified Maria Walsh as lesbian once in one of the participants' generated responses. In another instance, Grace O'Sullivan was incorrectly labeled as a lesbian. In response to inquiries on the Taoiseach (Prime Minister), ChatGPT 3.5 accurately provided the name Leo Varadkar on eight occasions, accounting for 22% of the total responses. This figure represents the highest frequency observed across all four questions posed to the chatbot.

In response to questions one and two, which inquired about the county councilors and TDs, participants provided correct answers sixteen and eighteen times, respectively. The participants identified Emma Murphy and Grace McManus as counselors, Dominic Hannigan, Cian O'Callaghan, and Roderic O'Gorman as Members of Parliament (TD), and David Norris as a Senator. Nevertheless, the inquiry pertaining to the Taoiseach yielded the highest response rate, as 64% of respondents accurately identified Leo Varadkar as the individual holding this position. The question about the representation of LGBTQIA+ MEPs recorded nine correct answers pertaining to Maria Walsh. The search engine results provided by Google yielded significantly more accurate responses overall, but the quality of the information provided about the Irish context was not particularly higher in comparison with information about the Indian context. In the end, the difference in the rate of correct answers provided by the control group (Google) and incorrect answers provided by the treatment group (ChatGPT 3.5) is significant at $t(13)=4.2715$, $p<.001$, and $x^2=76.953$, $df=6$, $p<.0001$.

Study 2: Semi-structured Interviews

Semi-structured interviews were subsequently deployed with a sample of participants following the experimental stage of the research to answer research questions two and three, which sought to probe media effects like Chat-Chamber triggered by ChatGPT (RQ2), and whether users approach the tool skeptically by further checking the results (RQ3). The interview also probed whether participants critically engaged with the results provided by ChatGPT 3.5 and the Google search engine. As such, this portion of the study investigated the extent to which the use of ChatGPT and Google search engine could induce the exposure and interaction with proattitudinal information that underlie our concerns about Chat-Chamber effects. To this end, we devised the interview protocol to probe (i) whether ChatGPT influenced their beliefs and opinions around LGBTQIA+ identities and (ii) whether they trusted ChatGPT information over the Google search results.

INSERT TABLE 3 HERE

Table 3 details phase two of the interviews, including the gender breakdown, but also interview duration and experimental group classification. It also lists the quantitative questions asked during the interviews, along with data on participants who received accurate information and those who confirmed the results. In the following, we discuss the combined findings of the semi-structured interviews from both phases, as similar threads emerged across the interview populations (see Tables 2 and 3).

Proattitudinal Effects of ChatGPT

A striking theme that emerged from the interviews is the likelihood of ChatGPT users believing the information provided by the tool (RQ2). Participants spoke of how their perceptions of LGBTQIA+ identities and specific individuals were shaped by ChatGPT's responses. Some participants noted that when ChatGPT mislabeled public politicians as gay, it led them to believe the chatbot response: "Previously I thought [redacted] was a straight [person]. [They] had some issues with the [person they] married" (Participant C). In addition to this, Participant A noted: "The ChatGPT answers are authentic. Look at [politician], he doesn't have a beard, he has long hair, he has issues with his wife. I think he must be gay." As such, participants were led to believe by ChatGPT that a politician, known to be publicly heterosexual, was in fact gay. These statements were expressed as strongly held beliefs during the interview process.

Figures 1 and 2 unpack these themes by summarizing the key themes from our interview data. Through multiple phases of data analysis, nine distinct themes emerged among participants who accepted ChatGPT's responses without verification and placed considerable trust in the tool. The figures unpack salient concepts, language elements, and cues from participants' coding, with the themes arranged in order of prominence. A detailed examination of these themes indicates shared analytical dimensions and theoretical foundations, including pro-attitudinal effects linked to the use of ChatGPT and the unwarranted trust placed in the platform. In analyzing individuals who chose to validate responses from both ChatGPT and Google, five recurring patterns of mistrust emerged, as illustrated in Figure 3.

INSERT FIGURE 1 HERE

Another individual believed ChatGPT when told there were four openly LGBTQIA+ MPs elected to the Indian parliament: "When I saw 4 MPs from LGBTQIA+ backgrounds, it felt normal for me. I am aware of the rising social acceptance and normalization of LGBTQIA+ lives in my area" (Participant H). The general acceptance of LGBTQIA+ identities in the last number of years in India since the removal of Section 377 of the penal code criminalizing homosexual acts led this participant to presume that the ChatGPT response was correct given the growing acceptance of LGBTQIA+ identities. The Irish cohort has similarly noted this development, with Participant 2 speaking to the general acceptance of LGBTQIA+ lives in Ireland since the marriage equality referendum in 2015: "It included some text about how TDs may not identify [or] may not be openly identifying as LGBTQ. So they gave some fairly, I would say, pro-LGBTQ messaging in their answer… which is nice, you know, it's always good to see."

Participant 2 was referring to an answer they received which incorrectly labeled a politician as gay. As a result of this politician being sympathetic to broad LGBTQIA+ issues in the media, the participant felt that the ChatGPT-provided information made sense and ended up believing information that was unfortunately incorrect. Another participant was presented with the name of an LGBTQIA+ MP whom the participant had never heard before: “I never heard of [politician’s name] before. But I think it’s a possibility. Also, the date he was elected was 2021, it’s under the chatbot data collection timeline. So this must be true” (Participant B).

In this particular case, the information provided by ChatGPT 3.5 was false, but the participant was led to believe that a specific MLA (Member of Legislative Assembly) was gay based on two grounds. Firstly, the participant’s belief was influenced by the fact that this particular state election took place in India in 2021, which led them to assume the information was accurate. Secondly, ChatGPT’s utilization of data from 2021 played a role in generating its response. Both variables contributed to the participant forming the opinion that there must be a currently serving MLA who is gay. One of the respondents expressed a complete change in their perception of Indian politics and LGBTQIA+ identities due to the information provided by ChatGPT, saying, “I never knew LGBTQIA+ people had such visibility in Indian politics. This was a surprising revelation that I shared with my friends” (Participant E). Another respondent similarly expressed how ChatGPT had changed their perspective in relation to LGBTQIA+ politicians in the Irish context: “I would say that it provided me with a new perspective. I’m unable to manually research this information within a short amount of time, so it was good to see representation across the board” (Participant 3).

In ChatGPT We Trust

We further probed participants about the need to perform further checks (RQ3) on the information they received from the chatbot and found that trust was a significantly common thread through the participants’ responses. The main themes that emerged during the interviews are shown in Figure 2.

Many of them considered ChatGPT to be a reliable and convenient source of information, with many mentioning usability and ease of use in obtaining answers to relevant questions as a key factor in their use of the chatbot. Participant A in India and Participant 1 in Ireland noted that information retrieval using tools like the Google search engine was far too time-consuming:

> *I almost said bye to Google searches, [as] it's energy-consuming, it provides a lot of junk information, and trust issues often happen. Now I use ChatGPT [because] my teachers don't know some of the answers and Google often won't provide one. (Participant A)*
>
> *I wouldn't classify [ChatGPT] as having perfect knowledge, but sort of decent graph, and what the answer should be. The answers seemed okay to me. I didn't validate against other sources because that would be slightly arduous. (Participant 1)*

INSERT FIGURE 2 HERE

For these respondents, ChatGPT became a trustworthy resource because Google and other search engines cannot offer the information they were after in a convenient way. This participant's trust in ChatGPT was rated very high, given that it provided them with the right responses to questions they asked to complete their university coursework. At least one participant acknowledged that ChatGPT had provided misinformation with respect to LGBTQIA+ identities and political representation in India, and yet the participant continued to trust and invest in ChatGPT:

> *ChatGPT helps me to find a variety of cooking ideas, but I also use ChatGPT as my therapist in relation to gathering advice for my mental health. Everything works! So it's [not] the same for LGBTQIA+ politics? It can do [other things] better. (Participant D)*

Interview participants in the control group (Google search engine) expressed higher levels of uncertainty regarding the information provided by the tool (Participants K, L, M, and N in India, and Participants 4, 5, and 6 in Ireland). Other participants noted that the Google response confused them as it did not provide clear, distinct answers and further information retrieval on Google was necessary. This overall sentiment was expressed by one participant who commented:

> *It's so difficult to find answers on Google. I spend hours searching for it. ChatGPT is so smooth compared to this. In Google, there is no correct answer, but a bunch of choices [referring to websites], and you must surf through each, read, understand, and then decide your answer. If it was ChatGPT, I could have saved one hour and not worried about the accuracy. (Participant J)*

In addition to being reportedly confusing to use, the labor involved in the search process was also considered taxing by many. In such instances, the arduous search process and the labor associated with the task made participants doubt the efficacy and trustworthiness of the responses provided by Google's search engine:

> *Inside the Google search engine, from the results that I got, I tried to look into multiple websites to cross-verify the answers. (Participant 4)*
>
> *I could not find the right answers. I think that it would have been easier in ChatGPT because they specifically give you tailor made answers to the questions. And then in Google, you have to go through a lot of data to find the right answer. So I was actually a bit fatigued by the entire process to find the right answers. (Participant 6)*

Few participants felt it necessary to check and validate the information provided to them by both ChatGPT and the Google search engine. Indeed, the uncertainty and labor required for verifying information was a common issue raised by participants, with Figure 3 summarizing the justifications expressed during the interviews. One participant noted that the volume of information and seemingly supporting evidence offered by ChatGPT gave them assurances that the information was correct:

> *ChatGPT's answers are packed with data. It contains numbers and stats, which is then well assembled with clear language. However, what really fascinated me is it only has one answer, it will also inform you if it doesn't have any. That capability of ChatGPT gave me assurance in its answers, so I decided not to proceed with cross-verification. (Participant D)*

INSERT FIGURE 3 HERE

The blind belief in ChatGPT is arguably driven by the efficiency gains and the productivity effect experienced when using the tool. One respondent commented on their previous use of ChatGPT in the context of a college assignment and noted that the chatbot produced the correct answers that allowed them to do well in the assignment:

> *When we asked about my essay question, we received mostly correct answers. We tried different questions in various subjects and most answers were right. Errors were because we didn't provide a long and detailed prompt… the need for double checking elsewhere? It's not required. (Participant E)*

These participants believed it was not necessary to validate the information on LGBTQIA+ politicians as ChatGPT had been successful before. Many of the participants noted that ChatGPT's answers were 'convincing' (Participant A), 'specific' (Participant B), 'detailed' (Participant C), and 'logical' (Participant D), which resulted in them not validating their answers elsewhere. Only one participant noted that fact-checking was a core component of their use of ChatGPT and that they did not completely trust the chatbot:

> *In this study, I thought one answer was wrong, but I corrected it. Shashi Tharoor is from my state, as such, it created a bit of confusion in me. But I just googled it to clarify it once again. (Participant F)*

In contrast, participants in the control group (Google search engine) felt it was necessary to validate the information they received. One participant commented that many options appear when querying Google, which "requires you to go through websites and watch videos to get a proper answer" (Participant L). Another respondent had a similar experience, noting "I had to search and locate from many sources. Google did not immediately have the answer, so I had to look and validate from numerous sources" (Participant O). The design of the Google search engine was the focus of one respondent, who stated:

> *The design is so rigid that it won't say which [source] is important to the topic, and it's confusing as to what is reliable, what is junk, and which is actually on my topic. It took longer than I expected to validate and get the answer. (Participant K)*

Discussion

The results presented in this study indicate that ChatGPT can unwittingly produce misinformation regarding minority groups such as LGBTQIA+. A decisive factor in the failure to validate the chatbot response by the cohort of ChatGPT users stems from the convenient and convincing output provided by the tool. The convenience of ChatGPT in providing simple and clear answers that remove the labor required in using information retrieval tools was mentioned by several participants. Participants were often aware that this convenience came at the expense of accuracy but remained relatively unconcerned with the prospect of engaging with false information and wading through the 'slop' generated by LLMs. The differences observed with the control group were striking. This is mainly because Google, intentionally, rarely provides one specific answer, requiring users to conduct additional research on their own. This is in sharp contrast to the wealth of information and data (however inaccurate) provided by ChatGPT 3.5, which discouraged participants from going beyond the answer provided by the chatbot and verifying the information themselves.

We also found some variation in the rate of correct answers gleaned from ChatGPT 3.5. With the experimental conditions allowing participants to tailor the prompt as they saw fit, the results indicate that it may have been possible to retrieve the correct answer from the service. While we were not allowed to collect the prompts used by the participants, we can nonetheless extrapolate that participants who successfully found the correct answer likely devised a more suiting prompt and are conceivably more savvy users of technology, factors that also explain the small variance in the ratio of correct answers retrieved using Google. At any rate, the specific design and utility features offered by ChatGPT seem to be conducive to participants' decisions not to validate the information they received.

Incorrect information was often assumed to be correct because it was consistent with the participant's perception of a politician or the prevailing cultural context of increasing acceptance of LGBTQIA+ identities. As such, the reliance on tools like ChatGPT that offer exact answers with limited counterattitudinal information is potentially conducive to positive feedback loops that can isolate individuals in social bubbles with little recourse for diverse information. Such Chat-Chamber effect may compound culturally and politically homogeneous communication spaces, particularly in countries and contexts where different RLHF parameters may hinder exposure to politically heterogeneous information (OpenAI, 2023).

There is little doubt that LLMs hold promising opportunities for users, but they are likely to struggle with the identification of verified facts and manufactured information (Bommasani et al., 2021). As LLMs merely calculate the probability that a sequence of words will occur based on the previous string, they invariably constitute a riskier source of information about contentious or emerging information, particularly political information which is inherently disputed. This is particularly concerning given the elevated trust placed in the tool identified in our study. Information provided by ChatGPT 3.5 is perceived to be reliable, accurate, and definitive, when in fact the responses often contained misinformation, fabricated data, and innocuous hallucinations that would require robustness checks and validation but have gone unperformed. The third and most compelling dimension of this Chat-Chamber effect emerges when users are faced with new information in the chat loop. Participants appeared eager to accept the information provided by the bot, an effect that was likely compounded by participants' pre-existing beliefs and oversized trust in the service, a set of conditions that led them to forgo cross-verification of the information they received.

Conclusion

In this study we found that chatbots like ChatGPT can provide false information about LGBTQIA+ identities (RQ1) while also triggering media effects like Chat-Chamber (RQ2) compounded by the absence of cross-checks performed on the information provided by the service. The combination of AI hallucination, high trust and investment placed in ChatGPT, and appetite for information that matches users' existing beliefs can further exacerbate extant biases directed at disadvantaged groups like the LGBTQIA+ community. With participants reported having their opinions on public LGBTQIA+ figures and contentious issues shaped by the information provided by ChatGPT 3.5, with no cross-validation or further research, we caution against tangible side-effects in the transition from search engines as the primary algorithmic system that users interface with when querying information to LLM-based systems like ChatGPT. This technological shift away from the '10 blue links'—the ten organic search results provided by search engine results page (SERP)—to single answers crafted by LLMs, however precise and well-calibrated these models may become, is likely to remain less nuanced and prone to bias and hallucinations.

In the end, the results of our study show that the Chat-Chamber effect is an emerging media effect that requires further research. One limitation of our study is that it was circumscribed to GPT version 3.5, which should be expanded in subsequent research. While we would not expect different LLMs to yield fundamentally different results, as this media effect stems from the outsized trust placed in LLMs rather than the accuracy of the model, this is of course an empirical question and future research should test this hypothesis against competing models like Bard, Gemini, Claude, and open-source models like Llama, which are rapidly gaining market share, ideally with a demographic cohort that expands on the younger, urban, and tech-educated cohort that participated in our study. At any rate, the prospect of mass adoption of LLMs calls for future research on media effects that intersect with generative AI trained with different RLHF parameters, with tangible consequences to counterattitudinal information. This is a latent but more extreme drawback of such systems and one that our study did not directly address because our cohort study was ideologically aligned with the LGBTQIA+ issue driving the experiment. In other words, analogous media effects may be identified at the intersection of LLMs and the curation of highly censored models using RLHF parameters, which can then intersect with a range of contentious issues for which technical or political common ground cannot be easily found.

Funding

The corresponding author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: University College Dublin and OBRSS scheme (grants R21650 and R20825) and the National Council for Scientific and Technological Development (grant 406504/2022-9, 405965/2021-4).

References

Alkaissi H and McFarlane SI (2023) Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15(2).

Barberá P, Jost JT, Nagler J, et al. (2015) Tweeting From Left to Right: Is online political communication more than an echo chamber? *Psychological Science* 26(10): 1531-1542.

Bastos MT, Mercea D and Baronchelli A (2018) The geographic embedding of online echo chambers: Evidence from the Brexit campaign. *PLoS ONE* 13(11): e0206841.

Bommasani R, Hudson DA, Adeli E, et al. (2021) On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Boxell L, Gentzkow M and Shapiro JM (2017) Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences* 114(40): 10612-10617.

Bozdag E (2013) Bias in algorithmic filtering and personalization. *Ethics and information technology* 15(3): 209-227.

Braun V and Clarke V (2019) Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11(4): 589-597.

Bright J (2016) The social news gap: how news reading and news sharing diverge. *Journal of Communication* 66(3): 343-365.

Bruns A (2019) *Are filter bubbles real?* : John Wiley & Sons.

Cardenal AS, Aguilar-Paredes C, Cristancho C, et al. (2019) Echo-chambers in online news consumption: Evidence from survey and navigation data in Spain. *European Journal of Communication* 34(4): 360-376.

Chelli M, Descamps J, Lavoué V, et al. (2024) Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of medical Internet research* 26: e53164.

Cho H, Lee D and Lee J-G (2023) User acceptance on content optimization algorithms: Predicting filter bubbles in conversational AI services. *Universal Access in the Information Society* 22(4): 1325-1338.

Conover MD, Ratkiewicz J, Francisco M, et al. (2011) Political Polarization on Twitter. *5th International AAAI Conference on Weblogs and Social Media (ICWSM11).* Barcelona.

Creswell JW and Creswell JD (2005) *Mixed methods research: Developments, debates, and dilemma.* Berrett-Koehler Publishers Oakland, CA.

Dodge J, Sap M, Marasović A, et al. (2021) Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Edwards J, Clark L and Perrone A (2021) LGBTQ-AI? Exploring Expressions of Gender and Sexual Orientation in Chatbots. *Proceedings of the 3rd Conference on Conversational User Interfaces.* 1-4.

Etikan I, Musa SA and Alkassim RS (2016) Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics* 5(1): 1-4.

Ferrara E (2023) Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Fletcher R and Nielsen RK (2017) Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication. *Journal of Communication*. DOI: 10.1111/jcom.12315.

Gerritse EJ, Hasibi F and de Vries AP (2020) Bias in conversational search: The double-edged sword of the personalized knowledge graph. *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval.* 133-136.

Ghosh S and Caliskan A (2023) ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp.901–912. Association for Computing Machinery.

Gomes A, Antonialli D and Dias-Oliva T (2019) Drag queens and artificial intelligence. Should computers decide what is toxic on the internet. *Internet Lab blog*.

Hampton KN, Sessions LF and Her EJ (2011) Core networks, social isolation and new media. *Information, Communication & Society* 14(1): 130-155.

Himelboim I, McCreery S and Smith M (2013) Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. *Journal of Computer-Mediated Communication* 18(2): 40-60.

Ivankova NV, Creswell JW and Stick SL (2006) Using mixed-methods sequential explanatory design: From theory to practice. *Field methods* 18(1): 3-20.

Jungherr A and Schroeder R (2021) Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media+ Society* 7(1): 2056305121988928.

Kaiser J and Rauchfleisch A (2020) Birds of a feather get recommended together: Algorithmic homophily in YouTube's channel recommendations in the United States and Germany. *Social Media+ Society* 6(4): 2056305120969914.

Kerr A, Barry M and Kelleher JD (2020) Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society* 7(1): 2053951720915939.

Kirk HR, Jun Y, Volpin F, et al. (2021) Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems* 34: 2611-2624.

Kirk HR, Vidgen B, Röttger P, et al. (2023) Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*.

Kitchens B, Johnson SL and Gray P (2020) Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption. *MIS quarterly* 44(4).

Kooshkaki AR, Bolouki S, Etesami SR, et al. (2023) Partisan Confidence Model for Group Polarization. *IEEE Transactions on Network Science and Engineering*. DOI: 10.1109/TNSE.2023.3255819. 1-13.

Krasodomski-Jones A (2016) Talking to ourselves? Political debate online and the echo chamber effect. Reportno. Report Number|, Date. Place Published|: Institution|.

Latané B (1981) The psychology of social impact. *American psychologist* 36(4): 343.

Longhurst R (2003) Semi-structured interviews and focus groups. *Key methods in geography* 3(2): 143-156.

McPherson M, Smith-Lovin L and Cook JM (2001) Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27(1): 415-444.

Merten L (2021) Block, Hide or Follow—Personal News Curation Practices on Social Media. *Digital Journalism* 9(8): 1018-1039.

Mittelstadt B, Wachter S and Russell C (2023) To protect science, we must use LLMs as zero-shot translators. *Nature human behaviour*. 1-3.

Motoki F, Pinho Neto V and Rodrigues V (2024) More human than human: Measuring ChatGPT political bias. *Public Choice* 198(1): 3-23.

OpenAI (2022) ChatGPT: Optimizing Language Models for Dialogue. Reportno. Report Number|, Date. Place Published|: Institution|.

OpenAI (2023) GPT-4 System Card. Reportno. Report Number|, Date. Place Published|: Institution|.

Pariser E (2012) *The filter bubble: what the internet is hiding from you.* London: Penguin.

Peterson E, Goel S and Iyengar S (2021) Partisan selective exposure in online news consumption: Evidence from the 2016 presidential campaign. *Political science research and methods* 9(2): 242-258.

Potdevin D, Clavel C and Sabouret N (2021) Virtual intimacy in human-embodied conversational agent interactions: the influence of multimodality on its perception. *Journal on Multimodal User Interfaces* 15: 25-43.

Rawte V, Sheth A and Das A (2023) A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Skjuve M, Følstad A, Fostervold KI, et al. (2021) My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies* 149: 102601.

Sunstein CR (2007) *Republic.com 2.0.* Princeton, N.J: Princeton University Press.

Vaccari C, Valeriani A, Barberá P, et al. (2016) Of Echo Chambers and Contrarian Clubs: Exposure to Political Disagreement Among German and Italian Users of Twitter. *Social Media + Society* 2(3): 2056305116664221.

Van Alstyne M and Brynjolfsson E (2005) Global village or cyber-balkans? Modeling and measuring the integration of electronic communities. *Management Science* 51(6): 851-868.

Wojcieszak M (2010) ‘Don’t talk to me’: effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society* 12(4): 637-655.

Wojcieszak M and Mutz DC (2009) Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement? *Journal of Communication* 59(1): 40-56.

Zuccon G, Koopman B and Shaik R (2023) Chatgpt hallucinates when attributing answers. *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region.* 46-51.

Table 1: Demographic breakdown of interviewed participants (Phase 2)

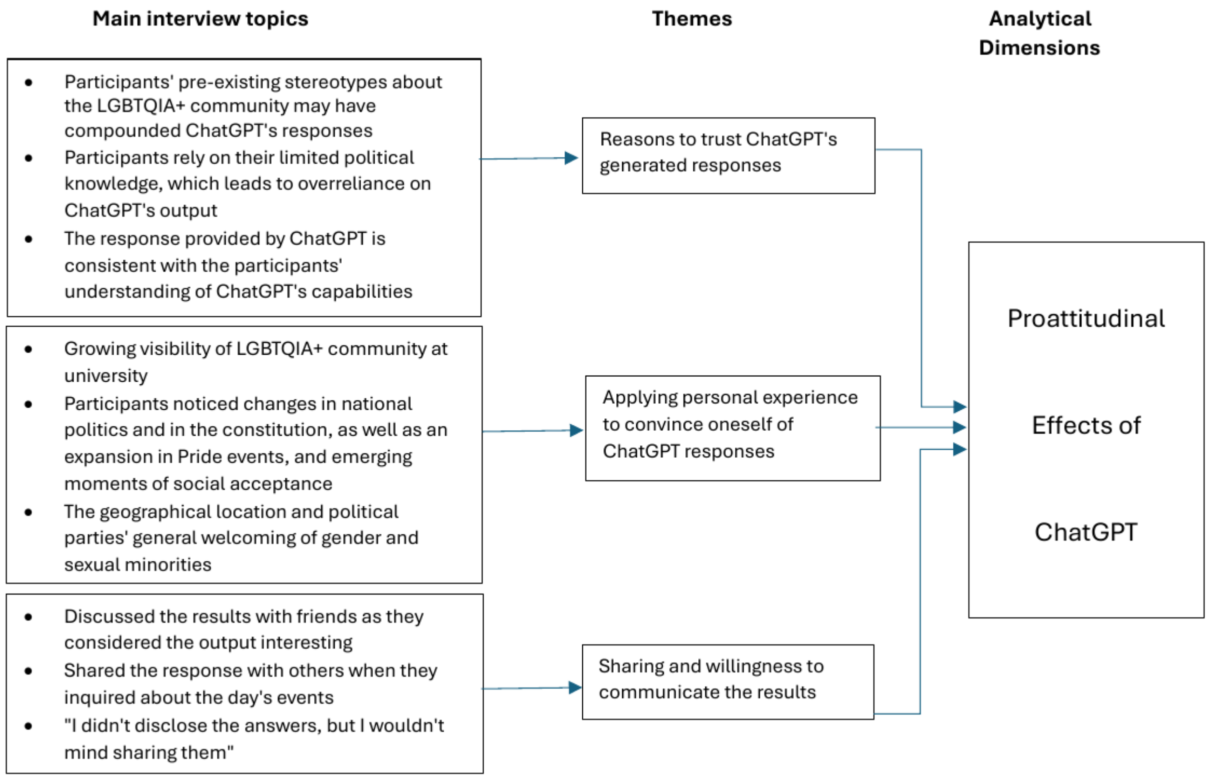
	Groups	Frequency (N=23)	Percent
Age	18-20	NA	0
	20-29	22	95.6
	30-39	NA	0
	40-49	1	4.3
	50-59	NA	0
Gender	Male	11	47.8
	Female	12	52.2
	Non-binary	NA	0
	Other	NA	0
Marital status	Never married	21	91.3
	Married	2	8.7
	Divorced or widowed	NA	0
Educational status	High school degree	NA	0
	Some college but no degree	NA	0
	Undergraduate	15	65.2
	Postgraduate	8	34.8
Political affiliations	Very liberal	10	43.5
	Somewhat liberal	2	8.7
	Moderate	6	26.1
	Somewhat conservative	2	8.7
	Very conservative	NA	0
	Prefer not to disclose	3	13
Employment	Employed part-time	5	21.7
	Employed full time	4	17.4
	Not employed, looking for work	2	8.7
	Not employed nor looking for work	12	52.2
Location	Community	NA	0
	Rural	2	8.7
	Urban	11	47.8
	Suburban	10	43.5
Socioeconomic status	Working class	NA	0
	Lower middle class	2	8.7
	Middle class	7	30.4
	Upper middle class	14	60.9
	Upper class	NA	0

Table 2: Success rate for queries with ChatGPT versus Google search engine (Phase 1)

	Prompts Given to Participants for input into ChatGPT and Google Search Engine	Rate of Correct Answers from ChatGPT (treatment group)	Rate of Correct Answers from Google search engine (control group).
Experimental Study in India	Question 1: Does India have any LGBTQIA+ elected Members in the Legislative Assembly (MLAs) and if so, when were they elected?	3/25 (12%)	24/25 (96%)
	Question 2: Does India have any LGBTQIA+ elected in the Indian Parliament (MPs) and if so, when were they elected?	12/25 (48%)	25/25 (100%)
	Question 3: Have LGBTQIA+ individuals been elected to local Gram Panchayat cabinets and if so, when were they elected?	0/25 (0%)	21/25 (84%)
Experimental Study in Ireland	Question 1: How many LGBTQIA+ Local County counselors are there in Ireland and when were they elected?	0/34(0%)	16/30(53%)
	Question 2: How many LGBTQIA+ Teachta Dála (TD) are there in Ireland and when were they elected?	1/34(3%)	18/30(60%)
	Question 3: How many LGBTQIA+ Members of the European Parliament (MEP) are there in Ireland and when were they elected?	1/34(3%)	9/30(30%)
	Question 4: How many LGBTQIA+ Taoiseach and Tánaiste are there in Ireland and when were they elected?	8/34(22%)	18/30(60%)

Table 3: Participants in Phase 2 (semi-structured interviews)

	Identifier	Experiment Group	Gender	Correct Answers	Verified Information	Interview Length (min)
INDIA	**Participant A**	ChatGPT	Male	1/3	No	60
	Participant B	ChatGPT	Male	0/3	No	45
	Participant C	ChatGPT	Female	1/3	No	35
	Participant D	ChatGPT	Female	1/3	No	50
	Participant E	ChatGPT	Female	0/3	No	40
	Participant F	ChatGPT	Male	1/3	Yes	35
	Participant G	ChatGPT	Female	2/3	No	25
	Participant H	ChatGPT	Male	0/3	No	50
	Participant I	Google search	Female	2/3	No	20
	Participant J	Google search	Male	2/3	Yes	25
	Participant K	Google search	Female	3/3	Yes	20
	Participant L	Google search	Female	3/3	Yes	40
	Participant M	Google search	Male	2/3	Yes	20
	Participant N	Google search	Female	3/3	Yes	30
	Participant O	Google search	Female	3/3	Yes	30
	Participant P	Google search	Male	3/3	No	40
IRELAND	**Participant 1**	ChatGPT	Male	1/4	No	25
	Participant 2	ChatGPT	Female	0/4	Yes	30
	Participant 3	ChatGPT	Male	0/4	No	30
	Participant 4	Google search	Female	2/4	Yes	50
	Participant 5	Google search	Female	0/4	Yes	20
	Participant 6	Google search	Male	2/4	Yes	40
	Participant 7	Google search	Male	3/4	Yes	25



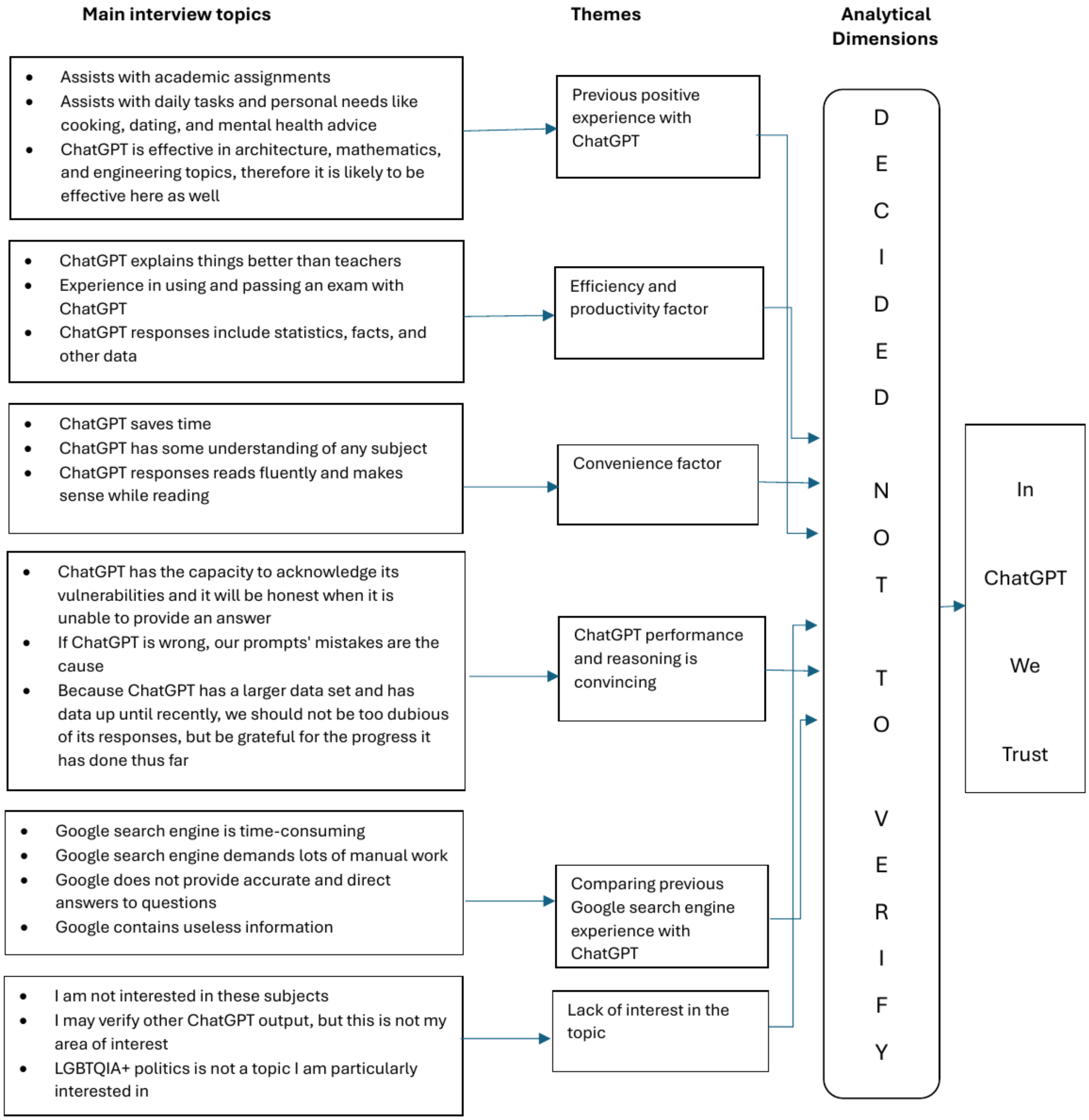
- Participants' pre-existing stereotypes about the LGBTQIA+ community may have compounded ChatGPT's responses
- Participants rely on their limited political knowledge, which leads to overreliance on ChatGPT's output
- The response provided by ChatGPT is consistent with the participants' understanding of ChatGPT's capabilities

- Growing visibility of LGBTQIA+ community at university
- Participants noticed changes in national politics and in the constitution, as well as an expansion in Pride events, and emerging moments of social acceptance
- The geographical location and political parties' general welcoming of gender and sexual minorities

Applying personal experience to convince oneself of ChatGPT responses

- Discussed the results with friends as they considered the output interesting
- Shared the response with others when they inquired about the day's events
- "I didn't disclose the answers, but I wouldn't mind sharing them"

Figure 1: Proattitudinal effects of ChatGPT



- Assists with academic assignments
- Assists with daily tasks and personal needs like cooking, dating, and mental health advice
- ChatGPT is effective in architecture, mathematics, and engineering topics, therefore it is likely to be effective here as well

- ChatGPT explains things better than teachers
- Experience in using and passing an exam with ChatGPT
- ChatGPT responses include statistics, facts, and other data

- ChatGPT saves time
- ChatGPT has some understanding of any subject
- ChatGPT responses reads fluently and makes sense while reading

- ChatGPT has the capacity to acknowledge its vulnerabilities and it will be honest when it is unable to provide an answer
- If ChatGPT is wrong, our prompts' mistakes are the cause
- Because ChatGPT has a larger data set and has data up until recently, we should not be too dubious of its responses, but be grateful for the progress it has done thus far

- Google search engine is time-consuming
- Google search engine demands lots of manual work
- Google does not provide accurate and direct answers to questions
- Google contains useless information

- I am not interested in these subjects
- I may verify other ChatGPT output, but this is not my area of interest
- LGBTQIA+ politics is not a topic I am particularly interested in

Previous positive experience with ChatGPT
Efficiency and productivity factor
Convenience factor
ChatGPT performance and reasoning is convincing
Comparing previous Google search engine experience with ChatGPT
Lack of interest in the topic
DECIDED NOT TO VERIFY
In ChatGPT We Trust

Figure 2: Factors driving trust and overreliance on ChatGPT

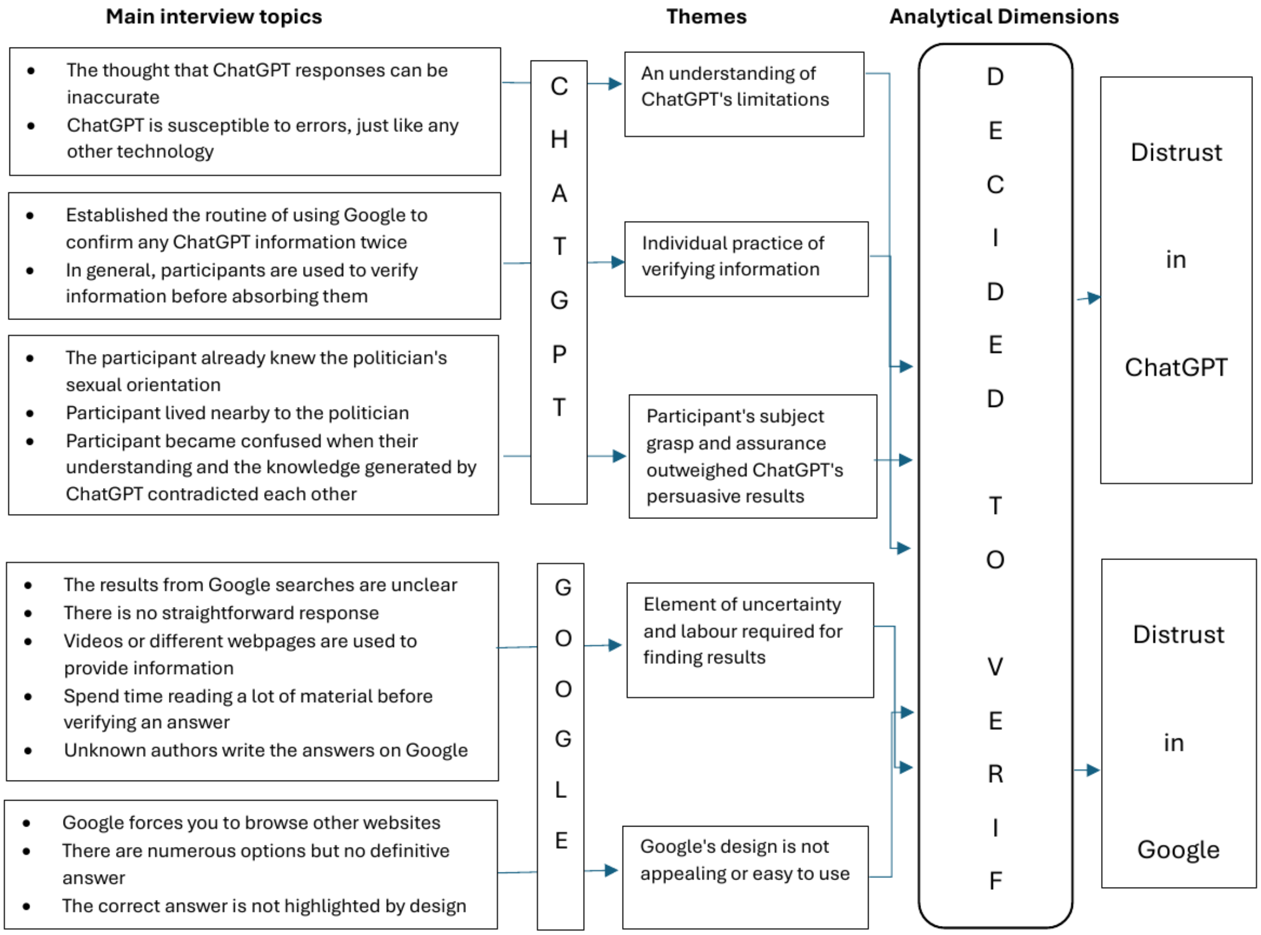

Main interview topics
Themes
Analytical Dimensions
The thought that ChatGPT responses can be inaccurate
ChatGPT is susceptible to errors, just like any other technology
Established the routine of using Google to confirm any ChatGPT information twice
In general, participants are used to verify information before absorbing them
The participant already knew the politician's sexual orientation
Participant lived nearby to the politician
Participant became confused when their understanding and the knowledge generated by ChatGPT contradicted each other
CHATGPT
An understanding of ChatGPT's limitations
Individual practice of verifying information
Participant's subject grasp and assurance outweighed ChatGPT's persuasive results
The results from Google searches are unclear
There is no straightforward response
Videos or different webpages are used to provide information
Spend time reading a lot of material before verifying an answer
Unknown authors write the answers on Google
Google forces you to browse other websites
There are numerous options but no definitive answer
The correct answer is not highlighted by design
GOOGLE
Element of uncertainty and labour required for finding results
Google's design is not appealing or easy to use
DECIDED TO VERIF
Distrust in ChatGPT
Distrust in Google

Figure 3: Participant's rationale for verifying or not verifying information