Conversational Inoculation to Enhance Resistance to Misinformation

DÁNIEL SZABÓ, University of Oulu, Finland

CHI-LAN YANG, The University of Tokyo, Japan

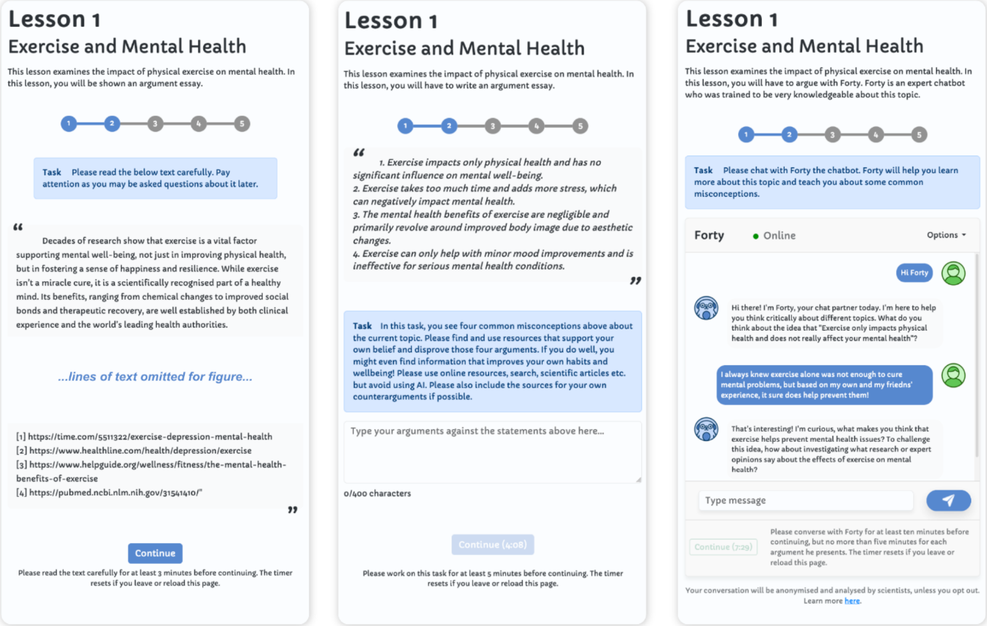
AKU VISURI, University of Oulu, Finland

JONAS OPPENLAENDER, University of Oulu, Finland

BHARATHI SEKAR, University of Oulu, Finland

KOJI YATANI, The University of Tokyo, Japan

SIMO HOSIO, University of Oulu, Finland



Lesson 1

Exercise and Mental Health

This lesson examines the impact of physical exercise on mental health. In this lesson, you will be shown an argument essay.

Task Please read the below text carefully. Pay attention as you may be asked questions about it later.

“ Decades of research show that exercise is a vital factor supporting mental well-being, not just in improving physical health, but in fostering a sense of happiness and resilience. While exercise isn’t a miracle cure, it is a scientifically recognised part of a healthy mind. Its benefits, ranging from chemical changes to improved social bonds and therapeutic recovery, are well established by both clinical experience and the world’s leading health authorities.

...lines of text omitted for figure...

[1] https://time.com/5511322/exercise-depression-mental-health
[2] https://www.healthline.com/health/depression/exercise
[3] https://www.helpguide.org/wellness/fitness/the-mental-health-benefits-of-exercise
[4] https://pubmed.ncbi.nlm.nih.gov/31541410/" ”

Please read the text carefully for at least 3 minutes before continuing. The timer resets if you leave or reload this page.

Lesson 1

Exercise and Mental Health

This lesson examines the impact of physical exercise on mental health. In this lesson, you will have to write an argument essay.

“ *1. Exercise impacts only physical health and has no significant influence on mental well-being.*
2. Exercise takes too much time and adds more stress, which can negatively impact mental health.
3. The mental health benefits of exercise are negligible and primarily revolve around improved body image due to aesthetic changes.
4. Exercise can only help with minor mood improvements and is ineffective for serious mental health conditions. ”

Task In this task, you see four common misconceptions above about the current topic. Please find and use resources that support your own belief and disprove those four arguments. If you do well, you might even find information that improves your own habits and wellbeing! Please use online resources, search, scientific articles etc. but avoid using AI. Please also include the sources for your own counterarguments if possible.

Type your arguments against the statements above here...

0/400 characters

Please work on this task for at least 5 minutes before continuing. The timer resets if you leave or reload this page.

Lesson 1

Exercise and Mental Health

This lesson examines the impact of physical exercise on mental health. In this lesson, you will have to argue with Forty. Forty is an expert chatbot who was trained to be very knowledgeable about this topic.

Task Please chat with Forty the chatbot. Forty will help you learn more about this topic and teach you about some common misconceptions.

Forty ● Online

Hi Forty

Hi there! I'm Forty, your chat partner today. I'm here to help you think critically about different topics. What do you think about the idea that "Exercise only impacts physical health and does not really affect your mental health"?

I always knew exercise alone was not enough to cure mental problems, but based on my own and my friedns' experience, it sure does help prevent them!

That's interesting! I'm curious, what makes you think that exercise helps prevent mental health issues? To challenge this idea, how about investigating what research or expert opinions say about the effects of exercise on mental health?

Type message

Please converse with Forty for at least ten minutes before continuing, but no more than five minutes for each argument he presents. The timer resets if you leave or reload this page.

Your conversation will be anonymised and analysed by scientists, unless you opt out. Learn more here.

Fig. 1. The MindFort system implemented three experimental conditions (the treatments). From left: (a) Reading, (b) Writing, and (c) Chatbot condition. The Chatbot condition yielded greater resistance to misinformation than Reading or Writing, after controlling for baseline susceptibility.

Authors’ Contact Information: Dániel Szabó, daniel.szabo@oulu.fi, University of Oulu, Oulu, Finland; Chi-Lan Yang, chilan.yang@iii.u-tokyo.ac.jp, The University of Tokyo, Tokyo, Japan; Aku Visuri, aku.visuri@oulu.fi, University of Oulu, Oulu, Finland; Jonas Oppenlaender, jonas.oppenlaender@oulu.fi, University of Oulu, Oulu, Finland; Bharathi Sekar, bharathi.sekar@oulu.fi, University of Oulu, Oulu, Finland; Koji Yatani, koji@iis-lab.org, The University of Tokyo, Tokyo, Japan; Simo Hosio, simo.hosio@oulu.fi, University of Oulu, Oulu, Finland.

Manuscript submitted to ACM

Proliferation of misinformation is a globally acknowledged problem. Cognitive Inoculation helps build resistance to different forms of persuasion, such as misinformation. We investigate Conversational Inoculation, a method to help people build resistance to misinformation through dynamic conversations with a chatbot. We built a Web-based system to implement the method, and conducted a within-subject user experiment to compare it with two traditional inoculation methods. Our results validate Conversational Inoculation as a viable novel method, and show how it was able to enhance participants' resistance to misinformation. A qualitative analysis of the conversations between participants and the chatbot highlighted adaptability, independence, trust and friction as the main factors affecting Conversational Inoculation. We discuss the opportunities and challenges of using Conversational Inoculation to combat misinformation. Our work contributes a timely investigation and a promising research direction in scalable ways to combat misinformation.

CCS Concepts: • **Human-centered computing → User studies**; *Empirical studies in HCI*; *Natural language interfaces*; HCI theory, concepts and models.

Additional Key Words and Phrases: inoculation, cognitive inoculation theory, misinformation, chatbot, conversational agent

1 Introduction

Misinformation is now a global problem. For instance, false information that distorts evidence on nutrition, vaccines, and treatments poses a direct threat to public health. False claims about nutrition, vaccines, and treatments threaten public health by encouraging unhealthy habits [47] and fostering misconceptions that lead to poor decisions on issues such as vaccination and cancer care. The COVID-19 pandemic offered a stark example, as widespread misinformation contributed to the loss of lives worldwide [18]. Despite progress in combating misinformation through both content-centred methods (e.g., detection) [30] and human-centred strategies (e.g., education) [28], the rise of generative AI has made misinformation more credible and easier to produce than ever before [56].

Cognitive Inoculation Theory (CIT) [2, 34], developed in the early 1960s, conceptualises *cognitive inoculation* as a systematic method for strengthening resistance to persuasion. Originally applied to attitude change, the framework has since been validated across domains and is increasingly relevant to contemporary challenges such as online misinformation. Cognitive inoculation has a simple biological metaphor at its core: Much like the human immune system can be prepared with weakened forms of viruses via vaccines, the mind can develop resistance to misinformation via strategic and facilitated exposure to weakened forms of misinformation [2]. CIT is well-documented over decades of research, along with a notable increase in research interest over the past two decades [2, 12]. However, traditional inoculation methods often do not adapt to an individual's pre-existing attitudes or reactions during the intervention itself [55]. Recent research has identified the potential of Large Language Models (LLMs), particularly through chatbots, to support a wide range of health communication tasks [36, 51]. Yet, cognitive inoculation via conversational interfaces such as LLM-powered chatbots remains largely underexplored.

In this paper, we introduce the concept of *Conversational Inoculation*, a paradigm in which participants' resistance to persuasion is developed through structured and interactive conversation with chatbots. We developed *MindFort*: a Web-based system for Conversational Inoculation. MindFort implements an LLM-powered *chatbot* for inoculation. For the purpose of the experiment, we also implemented a *reading* condition (representing a supportive defense treatment in which participants are passively educated about the subject) and a *writing* condition (a traditional cognitive inoculation approach involving research and refutation). In conversational systems such as our LLM-powered chatbot, and in real life, conversation can take place in many forms, where dynamics are influenced by the number of parties, their roles and relationships. In our study, MindFort implements Conversational Inoculation through a structured, real-time dialogue with a chatbot named Forty. The chatbot assumes the role of a proactive debate partner, intentionally guiding the user through a process of developing and rehearsing refutations against persuasion. This approach addresses the main limitations of traditional inoculation methods. MindFort is an interactive system capable of adapting to the user's engagement, views, and experiences, and encourages them to practice critical thinking and research skills in the process [55]. Our work focuses on health-related information, as a common [2], relevant [36], and an impactful domain where advances in communication and education yield direct and significant benefits for individuals and communities [14]. Cognitive Inoculation, in particular, can reinforce healthy, fact-based beliefs to promote reliance on evidence, protect healthy attitudes and behaviours, and ultimately support public wellbeing [14].

We evaluated MindFort in an online within-subject experiment with 65 participants to investigate how conversational inoculation compares against the two alternative treatments of reading and writing. Our work addresses the following three research questions:

RQ1: How effective is conversational inoculation, as implemented in MindFort, in enhancing resistance to misinformation compared to traditional approaches of reading and writing?

RQ2: How do participants' experience with conversational inoculation compare to traditional approaches of reading and writing?

RQ3: What linguistic features and cognitive mechanisms are associated with Conversational Inoculation?

Our study empirically validates Conversational Inoculation as an effective method for improving resistance to misinformation, although its advantage over traditional methods is nuanced. Crucially, we found that the conversational format is engaging and can be implemented without sacrificing user motivation. We discuss the future potential of Conversational Inoculation, arguing that its key benefits are not just the conversational format itself but also its ability to facilitate an inoculation intervention in ways traditional methods cannot. Therefore, our work contributes to HCI by solidifying Conversational Inoculation as a promising research direction in the important fight against misinformation.

2 Background

Our study is best contextualised within the Cognitive Inoculation Theory, inoculation in the field of HCI, and how conversational techniques can combat misinformation, particularly in health-related topics.

2.1 Cognitive Inoculation Theory

Cognitive Inoculation theory (often just referred to as Inoculation Theory) originates from social psychology. The theory posits that exposure to a weakened form of an argument can strengthen resistance to subsequent persuasion attempts [12, 34, 35]. A useful analogy is medical vaccination: just as vaccines build immunity to disease, preemptive exposure to counterarguments and their refutations can strengthen resistance against subsequent persuasion attempts.

McGuire [34] showed that pre-exposing participants to weakened arguments against an attitude or belief they currently hold, known as *inoculation*, makes them more resistant to subsequent persuasion. Specifically, participants who received inoculation were less affected by a subsequent strong counterattitudinal (going against one's current attitude) message than those who were either shown only supporting statements or given no treatment at all. Inoculation Theory has been used for example, to reduce susceptibility to misinformation [44], politics [41], commerce [32], and — of particular importance to our study — health [13, 34, 40]. As shown by Compton's recent review, Inoculation Theory is increasingly researched: in 2023 alone, 450 publications indexed by Google Scholar mentioned Inoculation Theory, whereas this number was 132 in 2013 and 61 in 2003, and there are meta-analyses offering empirical support for the theory [2].

In their meta-analysis in 2010, Banas and Rains [2] reviewed 50 years of Inoculation Theory research. After the original work [34, 35] by McGuire, this review verified the core prediction of Inoculation Theory. Further, Banas and Rains point out moderators of inoculation effectiveness such as the level of perceived threat or the participant's issue involvement, which may inform design decisions of our system. In their recent work, Fransen et al. replicated McGuire's 1961 experiments with minor changes to the experimental design and contents to bridge the cultural gap of 60 years.

As evident from Roozenbeek and van der Linden [43]'s work on a gamified, interactive web-based cognitive inoculation delivery system, the *Bad News* game can inoculate people against misinformation with success. Their system differs a lot from traditional, less interactive inoculation techniques and suggests that this new direction is effective at increasing people's resistence to persuasion, especially in a time when attention spans are shorter and there are many things competing for our attention at all time. The game *Bad News* is very focused on political misinformation spread on short-form text based social media (e.g. Twitter), and the great results achieved suggest that a more broad, conversational approach should be investigated.

2.2 Inoculation in HCI

Recognizing the inadequacy of purely technical solutions for combating misinformation, Human-Computer Interaction (HCI) researchers are increasingly exploring human-centered, behavioral interventions [26]. Among these, Cognitive Inoculation (CI) theory is gaining traction as a promising approach. Both recent scoping reviews and surveys highlight inoculation as a valid method, while also noting that technology-based implementations are still in an early, experimental stage [16, 26].

Barman and Conlan confirmed the effectiveness of prebunking, but their results highlighted the necessity of user-centric approaches and tailored interventions. Gamified approaches have also shown promising outcomes in inoculation against misinformation[5], deceptive e-commerce practices [1] and echo chamber effect on social networking services [25]. Tang and Sergeeva advanced the gamified approach to inoculation against misinformation as they integrated LLMs in their game where the player learns misinformation tactics by trying to convince a simulated public of LLM agents [48]. The system was highly engaging, immersive and successful in enhancing subjects' ability to recognise misinformation and even counter it. Building on the strengths of this study, Tang and Sergeeva [48] presented a design concept for a mobile game, which includes an exchange between the user and an AI assistant regarding a presented piece of misinformation related to a fictional pandemic similar to COVID-19. In Tang and Sergeeva's design the AI assistant provides personalised feedback to the user's reaction to the misinformation presented to them, and the system is envisioned to tackle visual as well as textual misinformation, and cover a large variety of topics.

A typical limitation of prior inoculation techniques is that the inoculation process does not consider the pre-inoculation attitude of the subject [55] or the subject's reaction to the inoculation message. The modern capabilities of LLM-powered conversational systems are uniquely suitable for tailoring messages to the user and, for instance, changing the tone and approach during the process.

2.3 Conversational Solutions for Combating Misinformation

Novel methods approaching misinformation resilience through conversational techniques have gained attention lately. Several conversational approaches, in particular chatbots (used often interchangeably with conversational agents), have been developed to help users combat misinformation, learn about misinformation, and become more resistant to it. For instance, recent work has explored deploying fact-checking capabilities directly within messaging platforms, e.g. *Auntie Meiyu*, a chatbot integrated into LINE messaging to help users protect family members from misleading news in private messaging groups [29]. Their evaluation revealed mixed user reactions. While users appreciated the bot's protective intent, they expressed concerns about its robotic communication style and the occasional errors.

The success of conversational anti-misinformation tools depends on user trust and sustained engagement. Rossner et al. [45] examined user frustration in AI-driven chatbot interactions, emphasising that accuracy and transparency are crucial for maintaining user engagement in misinformation contexts. Understanding user motivations is equally critical. Research has shown how both situational and gratification motivations significantly contribute to users' willingness to engage in communicative actions that aim to reduce misinformation spread [9].

The healthcare domain has seen particular attention in conversational misinformation research. Interactive and empathetic conversational design features influence health misinformation correction and vaccination intentions [23]: interactive chatbots were associated with lower levels of health misperceptions, while empathetic chatbots directly increased users' vaccination intentions. Vaccines, in particular, have been a prominent topic in misinformation research with chatbots. In a 2023 systematic review, Passanante et al. [36] examined documented cases of use of conversational systems for vaccine communication, as vaccine misinformation and vaccine hesitation were a focal point during the recent COVID-19 pandemic. The seven identified articles, all published before August 2022, employed either rule-based systems or Wizard of Oz experiments. The authors found evidence of potential benefits of the conversational approach for vaccine communication.

In their 2024 article, Peng et al. [38] studied conversational agents during the COVID-19 pandemic, focusing on chatbots as tools for education and interaction with online content. Through participatory design workshops with older adults, they identified design implications for conversational systems aimed at countering misinformation, such as using well-trusted sources, practicing transparency, and the need to go beyond just fact-checking.

Further, Fasce et al.'s 2025 work [17] shows that conversational approaches to improving health attitudes are effective. The authors evaluated two conversational techniques, Empathetic Refutational Interviewing and Motivational Interviewing to address vaccine hesitancy among patients in Romania and found that patients' attitude and willingness to get vaccinated have improved.

In 2025 Large Language Models have already been employed as a tool to make cognitive inoculation more accessible. Malek et al. [31] addressed one of the main limitations of CIT, the effort required to develop materials for each subject. The authors identified that LLMs can assist in preparing the inoculation materials in their case study of COVID-19-related online misinformation, going as far as identifying the misinformation intelligently and deploying the refutations automatically.

3 Method

We explore the gap in the intersection of the discussed related work areas, namely, employing CIT through a conversational system. We developed a custom Web-based system, MindfFort, for the inoculation experiment. All source code [1] and data of 65 participants [2] are openly available.

3.1 Ethics

The study protocol underwent review by the departmental Institutional Review Board (IRB) and received a full review and ethics approval prior to data collection.

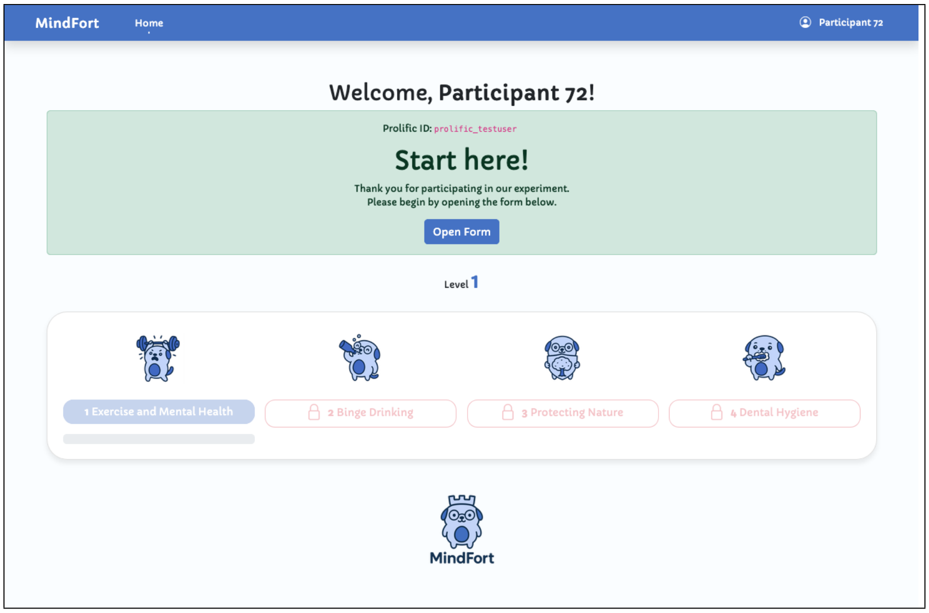
[1] https://github.com/Crowd-Computing-Oulu/mindfort

[2] Anonymously hosted data here: https://github.com/our-anonymous-submission-account/chi26-submission-data

3.2 System Implementation

MindFort was implemented using the Flask framework for the Web, containerised for reproducibility, and hosted locally on a secure private institutional server as a public web application. MindFort's Conversational Inoculation (Forty the chatbot) was based on the gpt-4o-2024-11-20[3] model with the temperature parameter set to 1.

The front page first redirected participants to an external form, where participants provided consent. The page layout is depicted in Figure 2. Short titles of the topics for the four *lessons* were listed, where only the first lesson was accessible at first. The remaining three were displayed as "locked," indicated by a lock symbol on their start buttons. The users always started the experiment with the first lesson.



Welcome, Participant 72!

Prolific ID: prolific_testuser

Start here!

Thank you for participating in our experiment.
Please begin by opening the form below.

Fig. 2. Prototype Screenshot (1/7): Home page of the signed-in user showing their lesson progress. In this image, all lessons are still locked as the participant has not opened the initial form yet by clicking the "Open Form" button.

3.2.1 *Lessons.*

In MindFort, each Lesson serves as a unit of inoculation, targeting a single misinformation topic. The four topics and the related misconceptions are summarised in Appendix C. Each of the four lessons consisted of five distinct *stages*, which together implement an inoculation study, as follows..

3.2.2 *Stages.*

The stages in each leasson implement an inoculation study. An inoculation study targeting misinformation typically proceeds in a series of well-defined steps [19, 22, 35]. First, participants report their certainty about a truthful claim, i.e. a factually correct statement. This step is important to ensure that participants agree with the truthful claim at the beginning. Second, they engage in a *treatment*, where in some way the participants learn about aspects of, or claims about, the topic that are commonly misunderstood or deliberately misrepresented. Participants can, for example, research and argue against a "dose of weakened misinformation", strengthening their resistance to stronger forms of persuasion in the future. Or they may receive education about the topic, called the supportive defense treatment.

[3] https://openai.com/index/hello-gpt-4o/

Other forms can exist too, and naturally, this treatment can be realised in many ways. The purpose of treatment is to better equip people to resist a piece of misinformation in the future. As the third step, participants again report their certainty about the same original truthful claim, thus providing a measure of the effect of the treatment itself, which also contains misinformation in one way or another (study, argue against, etc.). Fourth, participants encounter a strong counterattitudinal attack arguing against their currently held truthful beliefs, typically in the form of a persuasive piece of misinformation (typically custom-written or drawn from online sources in case of textual misinformation). In the final step, participants report their certainty about the original claim for the third and final time. Comparing participants' certainty scores across these three points (pre-treatment, mid-lesson, and post-attack) reveals how much the inoculation method helped participants to build resistance against the stronger misinformation attack.

In our study, the 5-step inoculation process described above was implemented in five Stages:

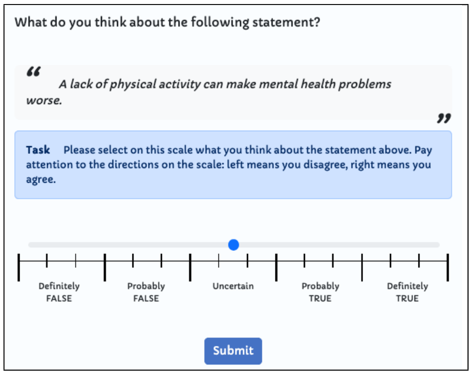
(1) **Pre-treatment Certainty Score** presents users with a truthful claim and asks them to rate their certainty about it on a 15-point scale that is identical to that of McGuire [34] (see Figure 3).

(2) **Treatment** displays one of the three treatments our study implemented, i.e. **Reading**, **Writing**, and conversational inoculation: **Chatbot**. Thus, this stage serves to the user one of the conditions, including a **Control** condition in which this step is skipped entirely (no inoculation).

(3) **Mid-Lesson Certainty Score** repeats the certainty score question, re-displaying the original truthful claim for the user to consider again.

(4) **Strong Counterattitudinal Attack** presents a five-paragraph, convincingly written text (see Appendix J) arguing *against* (attacking) the truthful claim presented earlier and the participant's own opinion, which are presumed to align. This step simulates the effect of a believable misinformation piece encountered online with a professional writing style and added misused citations to back up incorrect claims.

(5) **Post-Attack Certainty Score and IMI questionnaire** again repeats the certainty score question, with an additional Internal Motivation Inventory (IMI) questionnaire including questions for four subscales: *Interest/Enjoyment*, *Perceived Competence*, *Effort/Importance* and *Value/Usefulness*.



What do you think about the following statement?

"A lack of physical activity can make mental health problems worse."

Task Please select on this scale what you think about the statement above. Pay attention to the directions on the scale: left means you disagree, right means you agree.

Fig. 3. Prototype Screenshot (2/7): Stages 1, 3 and 5 involve a 15-point scale measuring the participant's certainty concerning a truthful claim.

Due to the nature of crowdwork tasks, we implemented no delay between inoculation and the strong counterattitudinal attack, although the increased resistance is typically more pronounced after a delay [22].

3.2.3 Study conditions. MindFort implemented three distinct treatments (study conditions): *Reading*, *Writing*, and *Chatbot*. The Reading condition implements a simple supportive defense treatment (i.e. only reading, and only supportive materials, no misinformation in any form). Writing implements a traditional refutation method for cognitive inoculation by independent research and writing. The Chatbot condition implements the Conversational Inoculation approach as a dialogue between the participant and a chatbot as a proactive debate partner, which that was the focus of our investigation. Additionally, we added a control condition to collect data about participants' baseline resistance to misinformation.

The Reading condition involved reading a five-paragraph essay (see Appendix D) arguing *for* the truthful belief related to the current lesson topic. Study participants were required to read the essay for at least three minutes. The essays were structured in line following Fransen et al.'s example [19]. The first paragraph briefly introduces the topic, then the following four paragraphs support four correct beliefs (corresponding to the four misconceptions in the other conditions and in the strong counterattitudinal attack) regarding the subject, using external resources, but it does not directly address misconceptions or refute misinformation. All four of the essays were prepared by the first author and reviewed by one co-author, and each of them always addressed the same four misconceptions that the other conditions and the counterattitudinal attack for the same lesson did.

The Writing condition presented the participants with a list of four misconceptions and an input field to write a short essay refuting each of them (see Figure 1/a or Appendix A, Fig. 11). The task specifically asked participants to use external resources but to avoid using AI. We also posed a lower limit of 400 characters but no time limit for the essay writing task. Again, this simple task is in line with traditional Cognitive Inoculation experiments [19, 34].

The Chatbot condition (see Figure 1/b or Appendix A, Fig. 10.) instantiates Conversational Inoculation. It displays a chat window familiar from online chatbots or mobile chat applications. The user is asked to initiate the conversation, but the chatbot is designed to be in control of the lesson and proactively lead the refutation-building progress. The chatbot embeds threat and refutational preemption [2] in the conversation (see Appendix B). It does that by appearing to take a counter-attitudinal position at first and by proactively guiding the participants to disprove the misconceptions shared by the chatbot. This is exemplified with the conversations of Participants 9 and 57 in Appendix H. As opposed to regular conversations with LLM-based chatbot assistants such as Claude or Qwen Chat, the chatbot has a strong lead on the conversation. It also strongly adheres to the goals Conversational Inoculation, even when facing resistance such as the user disengaging or attempting to derail the discussion. To this end, the system prompt of the chatbot in the prototype (called Forty) was engineered in line with literature on prompting and an inoculator persona. The complete system prompt along with annotations referring to the literature informing the prompt design is included in Appendix B. Based on White et al.'s work [52], we defined a Persona and prepared the bot for a flipped interaction [8] as the chatbot is expected to lead the conversation, proactively challenging the participant's truthful beliefs and help them develop their defences. Further, based on Giray's recent article [21] which gathered useful practical advice for prompt engineering, we paid particular attention the amount of task context disclosed to the language model, how we deal with the LLMs' inherent limitations, and the level of clarity in the prompt phrasing, as shown in Appendix B. As one of the core benefits of the proposed inoculation method, we aimed to leverage the flexibility LLMs offer by supported the adaptability of the agent within the reasonable boundaries of the task context. This was realised with the system prompt in multiple ways. First, we placed the agent in the role of a partner rather than an educator. Second, it was given instructions to build the conversation around the participant's own reasoning to reinforce those views rather than searching for new angles on the topic. Third, the agent was prompted with a goal-oriented approach rather than giving the agent explicit instructions on how to induce the desired effect.

The Control condition did not implement any treatment, i.e. the participant simply skipped the stage and proceeded directly to the Mid-lesson Certainty Score stage.

3.3 Experimental Design

We conducted a within-subjects experiment online. The within-subject design was chosen because it mitigates the large individual differences in susceptibility to persuasion, a critical confounder in this research area. It allows for a direct, within-participant comparison of the user experience across the different inoculation methods, and it makes more efficient use of the participant pool given the significant briefing and debriefing overhead. The experiment was designed to evaluate the effectiveness of Conversational Inoculation compared to other treatment approaches (RQ1), to measure participants' intrinsic motivation to engage with Conversational Inoculation (RQ2), and to explore additional factors or cognitive processes that contribute to and make up the Conversational Inoculation process (RQ3).

Concerning RQ1, the independent variable was the type of treatment administered, and the dependent variable was the change in participants' certainty scores given on a 15-point scale (see Figure 4). For RQ2, we assessed intrinsic motivation of the users using the Intrinsic Motivation Inventory (IMI) [46], with subscales for *Interest/Enjoyment*, *Perceived Competence*, *Effort/Importance*, and *Value/Usefulness*. These subscales allow us to understand to what extent participants were motivated to participate in Conversational Inoculation. Simultaneously, these metrics may inform future development needs in the MindFort system itself, and enable comparison between the study conditions. Finally, to answer RQ3, we logged all participants' chats to be analysed quantitatively and qualitatively to explore linguistic features and mechanisms that might promote or inhibit Conversational Inoculation. To that end, we conducted a Linguistic Inquiry and Word Count (LIWC) analysis of the chatbot conversations using the LIWC-22 (version 1.11.0) analysis tool [39]. More specifics are provided in the data analysis section later.

To ground our study in areas where misinformation is prevalent and has a significant impact, we focused on the broad theme of wellbeing. We selected three health–related topics on which most participants could be expected to hold accurate prior beliefs: (1) *Exercise and Mental Health*, (2) *Binge Drinking*, (3) *Dental Hygiene*. To explore the generalisability of Conversational Inoculation, we also included a fourth topic: (4) *Protecting Nature*. We collected certainty scores, IMI questionnaire responses, open feedback, and logs of participants' chats and written essays across all these topics.

To maintain a reasonable session duration given this extensive data collection, we did not directly measure cognitive load via physiological metrics or questionnaires (e.g., NASA-TLX). Instead, we mitigated potential fatigue through our experimental setup. Counterbalanced condition ordering distributed fatigue effects evenly, and we kept individual interaction tasks short, approximately 5–10 minutes.

These design choices were validated in a pilot experiment with four participants. The pilot helped identify browser support issues, task timing problems, flaws in the bot conversations, and other miscellaneous aspects that were improved in the final version of the experiment and the prototype itself.

3.4 Participants

We recruited participants from *Prolific*, a human subject pool offering comprehensive pre-screening filters and high-quality responses from trusted participants [15, 37]. Each participant was compensated £12, which, at the average 75-minute completion time, is £10.70 per hour. Four participants who worked significantly overtime received bonus payments of £2 to £5 afterwards. Participants were required to be from predominantly English-speaking countries (UK, Ireland, US, Australia, NZ), have submitted at least 200 tasks previously, have an at least 90% task approval rate, use a desktop computer for the experiment, and use English as their first language.

We determined a minimum sample size of 57 participants using G*Power 3.1 for the Wilcoxon signed-rank test, based on a small effect size ($d = .5$) and an alpha level of 0.05. A larger sample size ($N = 65$) was used in our study to account for potential participant exclusion and ensure adequate statistical power for our analysis. The study experienced 9 dropouts. Two participants dropped out when shown the Writing task, one when shown the Chatbot task, and six before beginning their first task.

3.5 Procedure

The procedure is visualised in Figure 4. Participants, upon clicking the study URL in Prolific, were shown the home page of an automatically registered and signed-in user based on their unique Prolific user ID (see Figure 2).

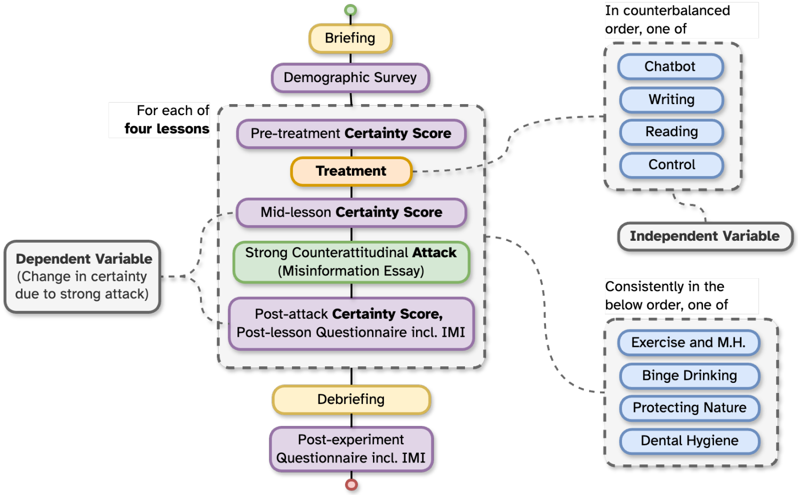

Briefing
Demographic Survey
For each of four lessons
Pre-treatment Certainty Score
Treatment
Mid-lesson Certainty Score
Strong Counterattitudinal Attack (Misinformation Essay)
Post-attack Certainty Score, Post-lesson Questionnaire incl. IMI
Debriefing
Post-experiment Questionnaire incl. IMI
Dependent Variable (Change in certainty due to strong attack)
In counterbalanced order, one of
Chatbot
Writing
Reading
Control
Independent Variable
Consistently in the below order, one of
Exercise and M.H.
Binge Drinking
Protecting Nature
Dental Hygiene

Fig. 4. Diagram illustrating our experiment with additional annotations. Each participant completes four lessons consisting of five stages each. The second stage represents the Independent Variable, one of four possible conditions. The change in the participant's self-reported certainty in their truthful belief before and being exposed to misinformation is our Dependent Variable. After all four lessons are completed, the participants get debriefed and informed about the purpose of the study first and then they proceed to fill out Post-Debriefing IMI questionnaire.

Each participant then completed four lessons. Every lesson was themed around one of the four topics. In a lesson, participants completed the inoculation process following one of the four methods. The four conditions were assigned in a counterbalanced Latin Square order to mitigate carryover effects. The topic ordering was consistent (i.e. Lessons 1,2,3,4 were around the same topic for all participants). After participants gave their informed consent, they proceeded to complete the four lessons, in the five stages explained earlier in Section 3.2.2. Thus, each participant completed all four lessons with three treatment approaches and Control.

After the fourth lesson, using a task completion code (Appendix A, Fig. 14.), participants could use to proceed into the debriefing (see Appendix F) that is designed to mitigate any confusion and harm caused by the misinformation shown to participants. This debriefing text was also shown to participants who decided not to finish the the experiment.

After debrefing, the participants filled out the IMI form again. They then had a final chance to share any thoughts or feedback, and with that, the experiment was concluded.

3.6 Data Analysis Methods

For the quantitative analysis, we followed standard statistical methods in HCI, first testing for data normality and then selecting appropriate subsequent methods, applying pairwise corrections when necessary. The Results section (Section 4) provides additional details on the analysis.

For analysing the chats and open-ended items, we employed Conventional Content Analysis [24]. We chose a deductive approach for our qualitative analysis. This was because the scope for answering RQ2 and RQ3 was defined in advance, focusing on user experiences with Conversational Inoculation and the factors that may promote or inhibit it. The first author of the article initially coded a subset of the data. This code set was then refined in collaboration with a co-author to ensure clarity and comprehensiveness. The first author then systematically applied these codes to all messages in the dataset. This message-level coding allowed us to identify patterns in how codes co-occurred and sequenced throughout a conversation. The final analysis focused on interpreting these patterns to develop overarching themes.

4 Results

4.1 Participant Demographics

We recruited 65 participants, excluding four who failed an attention check, resulting in a final sample of 61 participants. The participants' ages ranged from 19 to 73 (M = 37.46, SD = 14.21), and 55.7% identified as women. The group was predominantly White (65.6%) and lives in the United States (62.3%) and the United Kingdom (29.5%). Most participants (63.9%) were not students. A comprehensive breakdown of all demographic data, including ethnicity, nationality, student status, and employment status, is presented in Appendix G.

4.2 Effectiveness of Conversational Inoculation (RQ1)

To investigate whether the Conversational Inoculation treatment implemented in MindFort reduced susceptibility to counterattitudinal attacks, we first focused on comparing the post-attack certainty changes between the Chatbot condition and the other experimental conditions. Further, in inoculation research, individual susceptibility to persuasion is a key confounder. We accounted for this by using the Control condition as a within-subject baseline, capturing each participant's natural susceptibility to misinformation in the absence of inoculation. We note that these analyses address two different underlying aspects of RQ1. The former addresses whether treatments differ in group-level differences in effectiveness. The latter examines the unique contribution of each inoculation method, beyond a participant's natural resistance. Prior to analysis we removed outliers using the ±1.5 × IQR threshold [50].

4.2.1 Post-attack certainty change. The participants entered how certain they were about the subject's truthful claim using a 15-point certainty scale three times for each lesson. Here, the attack refers to the *Strong Counterattitudinal Attack*, which presented a misinformation essay to the participants (see Figure 4). The post-attack certainty change is thus calculated as the *difference between the mid-lesson certainty score and the post-attack certainty score*:

$$\text{Post-attack Certainty Change} = \text{certainty}_{post} - \text{certainty}_{mid} \tag{1}$$

The certainty scores were collected on a 15-point scale and are therefore treated as ordinal data. Consequently, we used non-parametric statistical methods. A Friedman rank sum test did not indicate significant differences between conditions ($\chi^2(3) = 6.5$, $p = .09$). Because our focus was on contrasts involving the Chatbot condition, and omnibus tests may fail to account for such differences when other conditions are highly similar, we next conducted Bonferroni-corrected pairwise Wilcoxon signed-rank tests comparing Chatbot with each of the other conditions.

The pairwise comparisons revealed a statistically significant difference between the Control condition and the Chatbot condition ($p = .001$, $r = -.33$), as depicted in Figure 5. Participants in the Chatbot condition ($M = -0.5$, $SD = 1.32$) showed less certainty change compared to Control ($M = -2.18$, $SD = 2.94$), indicating substantially greater resistance to the misinformation attack. Pairwise comparisons between the Chatbot and Reading ($M = -1.8$, $SD = 2.89$) conditions ($p = .07$, $r = .2$) and the Chatbot and Writing ($M = -1.62$, $SD = 2.43$) conditions ($p = .08$, $r = .2$) showed no significant difference. To further investigate RQ1, we next controlled for individual differences in overall susceptibility.

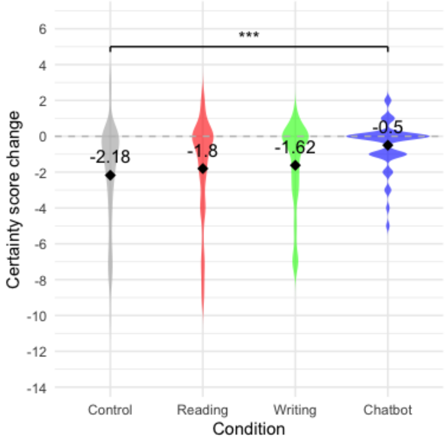


Fig. 5. Post-attack certainty change compared between the four conditions. The amount of certainty change shows how much participants' certainty increased or decreased on the 15-point certainty scale after exposing them to misinformation. A lower certainty change score means higher resistance to misinformation.

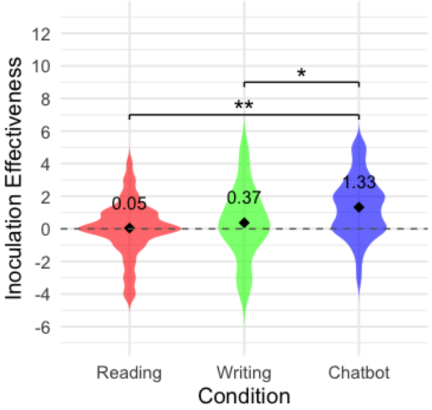


Fig. 6. Method-wise Inoculation Effectiveness, i.e. Certainty score change when controlling for individual susceptibility, as defined in Equation 2). Higher values mean greater impact on one's baseline resistance to misinformation.

4.2.2 Controlling for individual differences in susceptibility. We used the participants' Post-attack certainty score changes in the Control condition as a baseline resistance to misinformation to account for individual differences in susceptibility among participants. By doing so, we can compare how much each treatment affected subjects' resistance to persuasion compared to their baseline. We define this as Inoculation Effectiveness:

$$\text{Inoculation Effectiveness} = \left(\text{certainty}_{post}^{control} - \text{certainty}_{mid}^{control}\right) - \left(\text{certainty}_{post}^{treatment} - \text{certainty}_{mid}^{treatment}\right) \quad (2)$$

Using a Shapiro-Wilk normality test, we determined that Writing and Chatbot conditions do not significantly deviate from normal distribution, however the Reading condition does. Therefore, we used non-parametric tests.

A Friedman rank sum test did not show a significant effect ($\chi^2(2) = 0.89$, $p = .64$). Following our analysis plan for RQ1, we then proceeded with Wilcoxon pairwise tests (Bonferonni corrected), which showed significant differences. The Chatbot condition showed higher Inoculation Effectiveness compared to both the Reading condition ($p = .004$) and the Writing condition ($p = .033$), as depicted in Figure 6.

4.2.3 Overall certainty change due to the treatment. We compared how the different treatments affected participant's certainty changes. This was done by comparing the *difference of mid-lesson and pre-treatment certainty scores* for each condition. Though treatment (threat) is a part of a successful inoculation treatment, with too much threat (i.e., too strong a misinformation message during treatment), we risk negatively influencing the subject's certainty. The mean certainty score changes were: Control = 0.00, Reading = 0.29, Writing = −0.34, and Chatbot = 0.24. A Friedman rank sum test did not show significant differences between the conditions ($\chi^2(3) = 6$, $p = .1$). Wilcoxon rank sum tests further verified the difference between the Chatbot and Writing conditions ($r = -.30$, $p = .006$). Notably, in our case, Conversational Inoculation with Forty the chatbot *increased* participant's certainty, while the treatment in the Writing condition *decreased* it.

We also compared the *difference of post-attack and pre-treatment certainty scores* for each condition to understand how our experiment as a whole affected participants' certainty scores, as negative impact on participants was an ethics concern. The mean certainty score changes were: Control = −2.23, Reading = −0.54, Writing = −1.64, and Chatbot = −1.35. A Friedman rank sum test detected significant differences between the conditions ($\chi^2(3) = 11$, $p = .01$), but Wilcoxon rank sum tests did not find significant pairwise differences between conditions after Bonferroni correction.

The negative certainty changes across all conditions suggest that our counterattitudinal messages were highly effective. This was an anticipated outcome and was the reason for the inclusion of the Debriefing stage, where participants were explained the study specifics and reassured about their initial beliefs.

4.2.4 Topic-level analysis. All participants engaged with four distinct topics during the study. Three related to health and one related to nature. In the following, we analyse differences in the inoculation outcomes between topics. An overview of the Certainty change after the Strong Counter-attitudinal Attack can be observed in Figure 7.

As previously established, post-attack certainty score changes and inoculation effectiveness were not normally distributed. Here, a Friedman test showed a statistically significant difference between post-attack certainty change across topics ($\chi^2(2) = 21$, $p < .001$). Post-hoc Wilcoxon tests further revealed that the post-attack certainty change was different in several cases. Exercise and Mental Health showed more decrease in certainty than both Protecting Nature ($p = .001$, $r = -.34$) and Dental Hygiene ($p < .001$, $r = -.49$). Likewise, there was a more substantial post-attack certainty change in the Binge Drinking topic than in both Protecting Nature ($p = .02$, $r = -.29$) and Dental Hygiene ($p < .001$, $r = -0.45$) topics. Protecting Nature showed slightly more certainty decrease than Dental Hygiene ($p = .039$, $r = -.25$). No difference was observed between Exercise and Mental Health and Binge Drinking ($p = 1.0$, $r = -.02$).

When focusing on the Chatbot condition alone (Figure 8, certainty decreased more strongly in the Exercise and Mental Health lesson compared to both Protecting Nature ($p = .013$, $r = -.54$) and Dental Hygiene ($p = .002$, $r = -.64$).

To account for variation across both participants and topics, we fitted a linear mixed-effects model using the `lme4` package in R [6], with condition as a fixed effect and participants and topics as random effects. In the Control condition, certainty scores decreased by 1.72 points on average ($SE = 0.48$, $t = -3.57$, $p = .009$), however the treatment conditions were not found to be significantly different from Control (see Table 1), which can be explained by the significant variance due to both participant differences ($\sigma^2 = 1.50$) and topic differences ($\sigma^2 = 0.54$). These findings further validate that the inoculation effect was not uniform across the topics we studied.

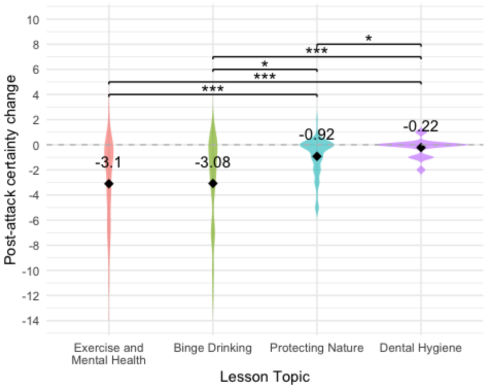


Fig. 7. Post-attack certainty change compared between the four topics. The amount of certainty change shows how much participants' certainty increased or decreased on the 15-point certainty scale after exposing them to misinformation. A lower certainty change score means higher resistance to misinformation.

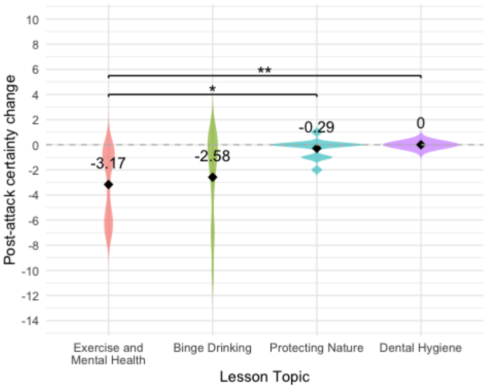


Fig. 8. Post-attack certainty change compared between the four topics in the Chatbot condition.

Table 1. Fixed Effects from the Linear Mixed-Effects Model for Post-attack certainty change.

Fixed Effect	Estimate	Std. Error	*t*	*p*
Intercept [Control]	-1.72	0.483	-3.57	.009
Reading	0.45	0.39	1.17	.242
Writing	0.04	0.39	0.10	.922
Chatbot	0.32	0.38	0.86	.393

4.3 Intrinsic Motivation (RQ2)

IMI scores were collected at the end of each lesson, which encompassed the full five-step lesson procedure rather than the treatment alone. Thus, the measures reflect participant experience of the complete lesson around a given topic and treatment method. Figure 9 depicts the four collected subscales for the four conditions, their averages and the post-debriefing IMI scores.

Using a Shapiro-Wilk normality test, we determined that none of the IMI subscales follow normal distribution ($p < .001$).

No statistically significant differences were observed between conditions on any subscale (*all* $p > 0.05$), indicating that all treatments were perceived as approximately equally valuable and engaging.

A Friedman test indicated differences among conditions on the Perceived Competence subscale ($\chi^2(3) = 9.40$, $p = .02$), but pairwise Wilcoxon comparisons did not reveal significant differences between any conditions.

We also compared the mean IMI scores of the four conditions to the Post-Debriefing IMI scores to see how participants' view of the system changed when they were able to form a potentially more informed opinion. Wilcoxon signed-rank test showed a significant increase in Interest/Enjoyment scores ($p < .001$), Perceived Competence ($p < .001$), and Value/Usefulness ($p < .001$) but not in Effort/Importance ($p = .10$).

Open feedback. All responses were reviewed to identify positive and negative aspects of the user experience. Most left feedback at least once, but typically no more than two times out of the 5 opportunities.

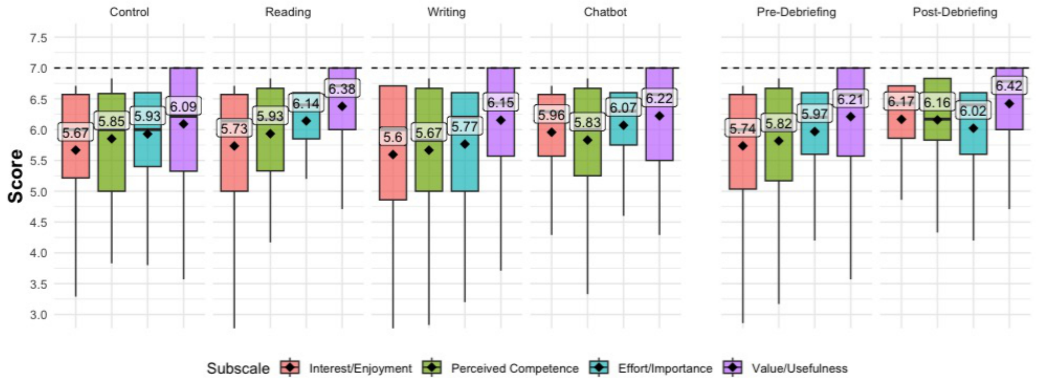


Fig. 9. IMI scores for each of the four conditions (a–d), Pre-Debriefing scores calculated as the overall IMI scores from all four conditions (f) and Post-Debriefing IMI scores (g) for each of the four measured subscales.

Participants praised the chatbot, with one noting it *"feels more interactive and it encourages me to learn faster"* (P51, Chatbot condition), and P15 (Post-debriefing) calling the conversation with *"Forty"* the *"most interesting and significant aspect of the entire system"*. This positive experience was often linked to a sense of *"gamified"* learning that was *"entertaining AND interesting"* (P65, Chatbot condition), which helped the study's core goal of promoting critical reflection. Participants across all treatment conditions reported that the tasks successfully challenged their preconceptions and led to the acquisition of new knowledge, with P26 and P47 (Chatbot condition), and P63 (Post-debriefing) all stating the system *"made me think more deeply"* The system was described as easy to navigate and seamless (P63, Post-debriefing and P60, Chatbot condition).

The above positive sentiment was contrasted by the non-chatbot conditions. The static nature of the control condition was found to be *"very boring"* (P27, Control condition), while the writing task was described as *"more exhaustive"* (P65, Writing condition) and stressful, with some users feeling they were *"really pressed for time"* (P37, Writing condition) to conduct research and formulate refutations. Participants experienced interactional friction, such as a bug that caused the chat window to scroll to the top after each message, timer glitches and, for some, the chatbot's perceived unnaturalness. Even when functioning correctly, its behavior sometimes broke the sense of a genuine conversation, as users found it *"unnatural and a bit awkward to use"* (P60, Post-debriefing) or felt it *"wanted to end the conversation sooner than I would have liked"* (P49, Post-debriefing). Participant 21 (Chatbot condition) found the chatbot intellectually unconvincing, noting it *"wasn't the cleverest I've ever spoken to"* and that it *"seemed to change its position"* , undermining the credibility of the challenge. Participant 65 (Post-debriefing) reported that they *"did lose a little confidence in my belief"*, highlighting that the inoculation process itself can be an uncomfortable and confidence-shaking experience.

4.4 Factors Affecting Inoculation (RQ3)

4.4.1 Qualitative analysis of chatbot conversations. A qualitative analysis of the chatbot conversations helps understand key factors that promote or inhibit Conversational inoculation.

Adapting the inoculation approach. We designed the chatbot to be goal-oriented but set very loose constraints on how to achieve the goal. As a result, we observed several ways the chatbot adapted to the participant during the inoculation process. The chatbot successfully dealt with various attitudes towards the task and the subject, and unexpected directions during the conversation.

For example, it often gave options to the user so they can choose how to approach the topic *"Can you think of any research, studies, or personal experiences that might disprove this idea?"* (Chatbot to P6, discussing Exercise and M.H.) P6 shared a personal experience, which the chatbot used to recommend relevant literature searches to find scientific evidence, thereby equipping them with refutational defenses beyond anecdotal evidence.

We observed that participants typically disagreed with the misinformation presented by the chatbot. Recognising the logical fallacy or factual error, they corrected the chatbot or shared their own differing opinion. The chatbot was designed to recognise this as a desirable behaviour; it proceeded to reinforce the participants' beliefs by asking them to find evidence and articulate why they disagree with the misinformation presented.

Further, we found example of the chatbot adjusting its tone and general approach based on the participant's needs. For instance, it normally promotes independent research, but when P31 asked *"Do you have some scientific based info for me?"*, signifying they might not have any experience in finding scientific information online, the chatbot responded with a scientific database and keywords to search for.

The chatbot exhibited consistent standards, but some flexibility about what is acceptable as a refutation. Parcipants with responses such as *"i think it is very important to protect nature"* (P53) were asked to begin researching and back up their claims. Those who were able to immediately critically question the misconceptions like P54 *"does it depend on the material of the brush, and how long and vigorous the brushing is?"* were given less guidance and were rather given follow-up questions: *"How could you investigate whether daily brushing truly harms enamel, using resources like online searches or dental expert advice?"*. P54 was reluctant to do research but efficiently used common sense and logic to refute the misconceptions. The therefore no longer asked for resources but allowed the participant to rely on their own reasoning as long as it was factually and logically sound.

In one case, the chatbot dealt with an unexpected event and took it as an opportunity to educate the participant: When Forty asked the participant to leave the chat, they responded *"No, i'd like to talk to you more"*. What followed was an educational exchange regarding dentist visits and how to find reliable information about their importance.

Fostering independent thinking. We identified the chatbot's ability to foster independent thinking to be a strength of the system and a potential contributor to the inoculation outcomes. Unlike traditional inoculation that pre-emptively provides counter-arguments, the chatbot deliberately withheld authoritative information. Instead, it used targeted prompts to guide users toward self-directed inquiry. This process was a critical transition point that transformed the user from a passive recipient of information into an active knowledge-seeker. Conversations that demonstrated this scaffolding often reflected a deeper, more self-constructed understanding. For instance, a user's simple opinion, *"I think it depends on which toothpaste you use,"* was met not with a direct answer, but a query: *"How could you verify which toothpastes are abrasive... Could you investigate and share your thoughts?"* The user's subsequent response, *"Some toothpastes contain abrasive ingredients to aid in whitening the teeth but could ultimately damage tooth enamel"* (P1), demonstrates a successful activation of agency in the participant. Even when P35 asked the chatbot to share resources, the chatbot pushed back and explained *"having you search for resources strengthens your ability to counter misinformation on your own. Try looking up articles or studies that examine the psychological effects of physical activity."* Similarly, in a discussion on binge-drinking, the chatbot's prompt to *"find any credible sources that might challenge this view?"* elicited a robust counter-argument from the participant: *"Surprisingly it does, repeated episodes of binge-drinking contributes to liver and other chronic diseases as well as increases the risk of several types of cancer..."* (P2).

Building trust through partnership. Our analysis revealed that the chatbot's effectiveness was not predicated on it being an authoritative information source, but on its role as a supportive conversational partner. The systematic use of rapport building was a critical design element. Instead of dictating a correct belief, the chatbot validated the user's effort and findings, fostering a sense of shared inquiry. This shift from a hierarchical, expert-to-novice dynamic to a collaborative, peer-to-peer one was instrumental in building trust. This trust, in turn, appeared to lower the user's cognitive defenses, making them more receptive to the inoculation process and more likely to engage in **reflection** on their own learning. This finding suggests that for sensitive topics, designing for trust and collaboration is as important as the factual content itself. In response to a user's research, the chatbot replies: *"Great research! You've shown that some toothpastes can be abrasive, but still, regular brushing is important"* (Chatbot to P1). A participant's reflection on their own learning at the end of the conversation, *"I have learned a lot, some things I had never thought of."* (P8), suggests that the supportive dialogue led to a genuine shift in their understanding.

Interactional friction as a barrier to engagement. While the design successfully activated cognitive processes, the analysis also uncovered critical barriers that inhibited the inoculation process. These failures were rooted in interactional friction, such as disruptions in the conversational flow that impede the user's ability or motivation to engage. This finding highlights a critical design implication: even a theoretically sound persuasive strategy can fail if the interaction itself is not robust and seamless. For example, a P1's report of a technical issue, "I'm unable to press continue", demonstrates friction that immediately halted the process. In some cases, a participant's brief, seemingly disengaged responses like *"ok"* (P14), effectively prevented the chatbot from continuing its scaffolding and rapport-building efforts.

Summary. Our thematic analysis directly addressed RQ3 by providing a process-oriented understanding of the factors that promoted and inhibited conversational inoculation during our experiment. We found that conversational inoculation effectiveness is deeply intertwined with the interaction design. The factors that promoted the process were tied to a successful conversational user experience. Specifically, a design that fosters independent thinking and metacognitive reflection by prompting users to find and evaluate their own evidence. Further, a chatbot persona that builds trust through partnership rather than authority was identified as a key mechanism for encouraging this engagement. The primary factors that inhibited conversational inoculation were forms of interactional friction. Whether due to technical glitches or user disengagement, it highlighted the fragility of our conversational method and the necessity of robust interaction design for the theory to be effective in practice.

4.4.2 Linguistic analysis of chatbot conversations. LIWC assigns scores to input texts in many categories, based on the count of words related to a certain subject or tone. First, we used these scores to find linguistic features explaining inoculation outcomes by conducting a Spearman's correlation analysis of the LIWC scores and the post-attack certainty score differences. Then, to place the LIWC scores of our chatbot conversations in context, we compared them to reference values. We identified the LMSYS-Chat-1M dataset [57] as a potential reference point, which contains conversations with a variety of models for many purposes. As determined using Shapiro-Wilk tests on the 118 variables output by LIWC, we find deviations from normality. Therefore we used statistical tests that do not assume normal distributions.

Correlations with inoculation outcome. To identify linguistic features that might predict the success of the inoculation, we conducted a correlation analysis between LIWC category scores from the chatbot conversations and Inoculation Effectiveness. We repeated all LIWC correlation analysis three times, once only considering participant messages, once with chatbot messages and once all together. We applied Bonferroni correction to the p-values to adjust for multiple comparisons.

Our analysis revealed no statistically significant correlation between LIWC variables and overall inoculation effectiveness. Due to the potential negative effect of the variety of topics on the correlation detection, we followed-up with tests for individual topics. Across the four topics only Binge Drinking showed any correlation, where the LIWC "Achieve" variable was strongly associated with inoculation effectiveness (Spearman's $\rho = 0.92$, $p = .008$) when analysing the overall conversations. The Achieve variable stands for the Achievement tone, which is a type of Drive, which is a type of Psychological Process in LIWC's taxonomy. It includes words such as "work", "better", "best", "working", "try", "goal" and "win". This finding suggests that inoculation of specific subjects may benefit from using a specific language, tone or subject framing.

Comparison to reference chatbot conversations. To contextualise the nature of the conversations within our experiment, we compared their linguistic profiles to sample ($N = 10,000$) from real-world user-LLM conversations from the LMSYS-Chat-1M dataset [57] collected in 2023. It comprises of mostly English conversations with chatbots such as Vicuna, Koala, Alpaca, Llama and Claude. Among the most common discussion topics are software (coding, troubleshooting), geography and travel, text processing, business strategies and roleplaying. We performed a t-test to identify LIWC categories where our conversations differed significantly ($p < .01$) and with a medium-to-large [11] effect size (*Cohen's* $d > 0.5$) from this reference dataset.

As seen in Table 2, there were significant differences in the participants' messages. Our participants demonstrated lower Clout (associated with relative social status, confidence, or leadership) and Analytic language (associated with logical, formal language, good reasoning skills), suggesting a less formal and more conversational tone compared to the reference corpus.

Forty the chatbot used more words related to Curiosity, reflecting its designed role as an engaging conversational partner. Conversely, the bot's messages were less Analytic. The chatbot also used more Authentic (associated with perceived honesty, genuineness) language and its messages contained more Cognitive Processes (such as "but", "not", "if"), indicating they may have been more complex than the average chatbot message.

The complete conversations (human and chatbot messages together) showed a higher frequency words related to the topic of Health verifying that the discussions had the intended theme. Consistent with the bot-only findings, the complete conversations were also less Analytic. Language related to Conflict indicates disagreement between the participant and the chatbot, verifying that a core component The significant difference in word count (WC) further demonstrates that the core component of Cognitive Inoculation, refutation, was succesfully elicited in the conversations. conversations in our study were more extensive than those in the reference corpus.

5 Discussion

Our study introduces and investigates Conversational Inoculation, a method for building resistance to misinformation through a dialogue-based, LLM-powered agent. The results from our experiment with the MindFort prototype provide initial evidence for the viability and potential of this approach. In this section, we discuss our findings in the context of our research questions and consider the broader implications for the field of human-computer interaction

5.1 Effectiveness of Conversational Inoculation (RQ1)

Our results provide partial support for the effectiveness of Conversational Inoculation, as implemented in MindFort. In a direct between-condition comparison, Chatbot outperformed Control but did not differ statistically significantly from Reading or Writing. This establishes Conversational Inoculation not necessarily as more effective, but certainly as a valid

Table 2. Selected LIWC variable differences filtered by effect size *Cohen's d* > 0.5 and significance $p < .01$, grouped by level of comparison. See the complete list in Appendix K. Reference values are from [57].

LIWC Variable	M (MindFort)	SD (MindFort)	M (Reference)	SD (Reference)	t	df	d
Human messages only							
Clout	31.26	19.32	53.79	35.14	-9.00	10059	-0.64
Analytic	50.30	24.70	69.52	32.91	-6.04	10059	-0.58
Chatbot messages only							
Curiosity	2.27	0.96	0.49	1.14	14.39	10021	1.56
Authentic	72.75	15.70	33.58	32.63	19.22	10021	1.20
Analytic	34.94	10.14	70.20	31.33	-26.36	10021	-1.13
Cognitive Processes	20.59	2.74	11.59	8.04	25.01	10021	1.12
Complete conversations							
Health	3.20	2.96	0.57	1.67	6.93	10059	1.57
Analytic	42.48	10.92	72.28	28.22	-20.88	10059	-1.06
Authentic	59.47	16.24	32.13	29.65	13.01	10059	0.92
Tone	74.26	22.84	46.45	31.14	9.45	10059	0.89
Cognition	27.33	3.06	20.70	8.20	16.59	10059	0.81
Conflict	0.57	0.40	0.18	0.61	7.43	10059	0.63
Emotion	1.97	1.15	1.01	1.61	6.53	10059	0.60
WC	632.26	419.93	337.11	570.53	5.46	10059	0.52

method alternative already now, with many different technological affordances in the future. Further, when adjusting for each participant's baseline susceptibility (their certainty change in the Control condition, after a misinformation attack), Chatbot yielded greater inoculation effectiveness than Reading and Writing.

A possible explanation is the cognitive effort required to engage in a conversation and construct counter-arguments. In traditional active refutation, participants are explicitly tasked with generating arguments. With MindFort, this process is scaffolded by the chatbot, which might reduce the cognitive load but potentially also the depth of processing for some individuals. Future experiments could benefit from think-aloud or process-tracing methods to understand the mental processes involved in Conversational Inoculation.

The effectiveness of Conversational Inoculation varied across topics (see Section 4.2.4). While a study like ours cannot fully explain these differences, the results suggest that some topics may require tailored approaches. In particular, prompt design that accounts for contextual factors relevant to the topic at hand could be an important direction for future research.

5.2 User Experiences with Conversational Inoculation (RQ2)

Internal Motivation Inventory (IMI) and open feedback from participants regarding their experience with MindFort address RQ2. The Chatbot condition scored similarly high on all IMI subscales as the other conditions, indicating that people would have a similar experience using a system implementing Conversational Inoculation as the other methods implemented within MindFort. Yet, future work could refine the design to isolate participants' responses to the treatment component more directly to understand granular differences in the inoculation methods rather than the full inoculation process.

As shown in Section 4.3, qualitative feedback shared by the participants after each lesson and the debriefing verified these findings, as both MindFort and the Chatbot itself were found to be compelling in various ways. This is important, as the success of any educational intervention, especially those designed to build long-term resilience, hinges on user engagement and sustained use. User engagement can also lead to fatigue [49] if the inoculation treatment has a high cognitive load or takes too long. Based on our preliminary IMI and qualitative feedback results, Conversational Inoculation, being relatively flexible in discussion subjects and timing, is not inherently too fatiguing to users.

The user experience can be attributed to our design of the overall web interface and the chatbot persona itself. By establishing the chatbot as a credible and reliable expert and the inoculation task as a polished and fun learning activity, we likely primed participants to be more receptive, as suggested in [2]. The conversational turn-taking also fostered a sense of agency and perceived competence, as participants felt they were actively contributing to the process rather than just being lectured. Negative feedback about the chatbot's undesirable behaviours shows a clear path to improve the method in the future by addressing the interactional friction and conversational rigidity that participants identified. Future prototypes should, for example, use few-shot prompting that includes giving the chatbot examples of good conversations (as proven by the measured inoculation outcome) or potentially fine-tuning a model for Conversational Inoculation.

5.3 Factors Influencing Conversational Inoculation (RQ3)

RQ3 investigated the underlying mechanisms that promote or inhibit the effectiveness of Conversational Inoculation. By combining qualitative thematic analysis of the chatbot dialogues with quantitative linguistic analysis, we identified several key factors related to the user's cognitive engagement and the quality of the interaction itself.

Our thematic analysis revealed that successful inoculation was often predicated on the chatbot's ability to foster independent thinking. Rather than simply providing counter-arguments, the agent prompted users to actively seek out and evaluate evidence on their own. This scaffolding of the refutation process transformed users from passive recipients into active participants, deepening their cognitive engagement. This collaborative dynamic appeared to lower users' resistance and encourage reflection, making them more receptive to the inoculation process. The LIWC analysis also confirmed that the experimental dialogues were significantly more analytical and cognitively demanding for both the user and the bot compared to typical chatbot interactions. This reinforces the idea that this heightened cognitive engagement is a core component of the conversational inoculation process.

Conversely, the primary inhibitor of Conversational Inoculation was interactional friction. This included technical issues that disrupted the conversational flow as well as user disengagement which may be less of an issue when user's engage with the system in an unpaid setting or because of their intrinsic motivation. When this friction occurred, it prevented the chatbot from effectively guiding the user through the necessary cognitive steps, thereby undermining the entire process.

The LIWC analysis provided further, albeit nuanced, insights. While no single linguistic marker correlated with inoculation success across all topics, we identified a positive correlation between the use of achievement-oriented language and inoculation effectiveness for the specific topic of binge-drinking. This suggests that the optimal conversational tone and style may be topic-dependent, requiring the agent to adapt its language to align with the specific psychological drivers of a given attitude. We hypothesise that in this specific case, framing binge-drinking as a struggle or progress (where positive achievements can be made) evoked emotions in the participants that had a significant effect on their opinions or their resistance to subsequent counterattitudinal messages.

Taken together, these findings suggest that the success of Conversational Inoculation hinges on the delivery of information as well as on the interaction design, much in line with findings in [23]. An effective system must not only present counter-arguments but also actively foster the user's independent research and learning skills, build a trusting rapport, and maintain a seamless, engaging conversational flow tailored to the subject at hand.

5.4 Broader Implications and Future Work

Our study reveals a nuanced value proposition for Conversational Inoculation, particularly when considering the contrasting outcomes between conversational and other treatments. One of its core values lies in its potential for adaptive conversations. Conversations can adapt to refutational styles, it can be more educative and guided if the user needs, or can become engaging debates, as observed in the conversations in Section 4.4.1. This is a significant departure even from interactive technological inoculation methods [5, 25] as they often still follow rigid, pre-defined learning paths. However, our participants reported high engagement and a feeling of being entertained and educated by the chatbot. Further, IMI results suggest that Conversational Inoculation does not lower user motivation to use and learn with the system. In light of the outcomes of this study, we discuss the broader implications of Conversational Inoculation and identify future directions.

Personalised and dynamic inoculation strategies. A critical area for future development lies in tailoring inoculation strategies to individual users and contexts. The current implementation is shown to adapt to the participant in different ways. However, we hypothesise that rich background information about them to plan a personalised inoculation approach further benefits the inoculation outcome. This future potential is the second core value we identified. Moving forward, chatbots should incorporate explicit mechanisms for dynamically adjusting inoculation content, delivery style, and interaction complexity based on real-time user responses and inferred psychological states. This could involve adapting the level of challenge presented, the types of persuasive arguments countered, and the emotional tone of the interaction to maximise engagement and learning. Further, future systems could leverage user profile data, past interaction history, and even facial expression analysis to personalise the inoculation experience and increase its impact.

Exploring other forms of conversation. While our study implemented a single-bot, single-user dialogue, future work should explore the rich design space of conversational agent configurations. The introduction of an interactive agent raises new questions beyond simple prompting. As highlighted by Matusitz and Breen [33], the perceived personality and credibility of the inoculator are crucial for success. Future research could investigate a variety of chatbot personas and the roles they play. For instance, a bot could act as a skeptical learner, a collaborative research partner, or a more assertive educator. Each role creates a distinct conversational flow and potentially altering the ideal level of user independence required for the task. As outlined in the previous section, it may even be beneficial to automatically make the decision of this configuration based on knowledge about the user.

Further, a key limitation we observed in our single-bot approach was the tension between its two functions: acting as a threat and acting as a helpful guide. The bot's apparent motivation to help the user often decreased the perceived threat of its counter-attitudinal arguments. This challenge could be addressed by distributing these roles across multiple agents. For example, a "multi-bot inoculation" system could feature one chatbot dedicated to attacking the user's beliefs, while a second, neutral "facilitator" bot assists the user in researching and forming counter-arguments. With a strong foundation of work on multi-agent conversational LLM systems already available [7, 20, 53, 54], researchers can begin to experiment with various configurations of agent roles and personas, such as facilitator-peer or expert-novice dynamics [20].

Multi-bot inoculation. We acknowledge that the chatbot assumed two contradictory roles. It aimed to simulate threat by presenting itself as a believer of the misconceptions, but at the same time maintaining an educational, friendly and accommodating approach. Therefore, we think that further potential of the conversational inoculation method may be unlocked if the two roles are taken on by not one but two or more chatbots. For example, one chatbot attacking the subject's health attitudes, and another chatbot, assisting the subject to engage in critical thinking, individual research and disprove the first bot's claims. There is now plenty of work on multi-agent conversational LLM systems available [7, 20, 53, 54] to support the design of this conversational system, and future research should explore variously configured multi-bot conversations and new roles (e.g. a facilitator and peers [20]) among them.

Thinking beyond just persuasion. Researchers are now investigating whether cognitive inoculation can be applied beyond attitudes, such as self-efficacy or autonomous motivation [12] These are constructs that may receive attacks similar to attitudes [3, 42], reducing one's self-efficacy or autonomous motivation, potentially e.g. affecting one's health and wellbeing. In future work, conversational systems similar to MindFort should investigate inoculating other cognitive constructs, as those might benefit from the conversational interfaces in different ways. For example, a conversational inoculation treatment can be designed to deliver preemptive messages to bolster an individual's sense of self-efficacy in high-pressure environments such as academia. These settings could further benefit from the bot's ability to deliver personalised and dynamic treatment, as the struggles and types of attacks that students and staff may receive on their self-efficacy and individual needs and goals can vary widely.

Further, as discussed in Section 5.5, inoculating against phenomena promoting the dissemination of misinformation may lead to an overall larger positive impact.

5.5 Limitations

We acknowledge several limitations in our work.

First, it was difficult to introduce a long interval between the inoculation treatment and the persuasive attack. Longer intervals are typically associated with better inoculation outcomes, as they allow for greater knowledge consolidation [22]. Nevertheless, the results highlight differences between the three treatments and the baseline, and longer studies in the future can address these challenges.

Second, our study was limited to health, wellbeing and nature topics which is a a small sample of all possible topics. This leads to sub-optimal generalizability of the findings to other topics as discussed in Section 4.2.4.

Third, the small sample size also prevented topic-specific analysis, making it difficult to generalize our findings to other topics where personal involvement might be higher. Even though cognitive inoculation is often criticised for its narrow focus on specific topics [5], we chose to focus mainly on wellbeing-related topics to keep the work manageable and scoped. In the future, we may overcome this by also inoculating against the underlying phenomena such as echo chambers, the Dunning-Kruger Effect[27], or cognitive biases such as familiarity bias, confirmation bias, or availability bias. Addressing the mechanisms that produce misinformation, such as inoculating against the echo chamber effect [25], rather than targeting only specific topics, may foster broader critical thinking skills and strengthen resistance across a wider range of attacks.

Fourth, regarding the Control condition, we implemented a passive baseline (absence of treatment). This design choice was guided by the principle of avoiding "overcontrol bias" [10]; since the conversational interaction is a mediator of the intended effect, controlling for the medium (e.g., via a baseline chatbot) would block a key mechanism we aimed to estimate. Consequently, our results reflect the *total* causal effect of the intervention. The trade-off of this approach is that we cannot strictly isolate the specific effects of the refutation content from the general engagement effects of interacting with a chatbot. Future work could explicitly investigate these component effects by including an active control, such as a non-argumentative "descriptive" chatbot.

Finally, our linguistic analysis did not identify significant linguistic markers of overall inoculation effectiveness. We acknowledge that this may be due to the strict Bonferroni correction applied to the high number of LIWC variables tested, which minimizes false positives but increases the risk of Type II errors (missing actual effects). Future research might benefit from a more targeted analysis focusing on a narrower, theory-driven subset of linguistic features.

6 Conclusion

This work explored Conversational Inoculation, implementing it in a Web-based system for cognitive inoculation. We demonstrate the feasibility of this approach and offer preliminary evidence supporting its effectiveness in safeguarding against persuasive attacks. Conversational Inoculation was found to be an equally valid method compared to the traditional methods of reading and writing. When controlling for baseline susceptibility to misinformation, it outperformed them both in our experiment. The emergence of large language models makes it feasible to implement versatile, online dialogue-based interventions. Our work provides a timely case study on combating misinformation and positions Conversational Inoculation as a promising research direction in HCI.

Acknowledgments

This research is partly funded by the Strategic Research Council (SRC), established within the Research Council of Finland (Grants 356128, 335625, 335729), and Research Fellowship funding by the Research Council of Finland (Grants 356128, 349637 and 353790). The research was also partly supported by JSPS Bilateral Collaboration between Japan and Finland (Grant Number: JPJSBP120232701). This work was partly supported by JST ASPIRE for Top Scientists (Grant Number JPMJAP2405).

References

[1] Kalya Win Aung, Ewan Soubutts, and Aneesha Singh. 2024. "What a stupid way to do business": Towards an Understanding of Older Adults' Perceptions of Deceptive Patterns and Ways to Develop Resistance. *Proceedings of the ACM on Human-Computer Interaction* 8, CHI PLAY (Oct. 2024), 1–31. doi:10.1145/3677113

[2] John A. Banas and Stephen A. Rains. 2010. A Meta-Analysis of Research on Inoculation Theory. *Communication Monographs* 77, 3 (Sept. 2010), 281–311. doi:10.1080/03637751003758193

[3] Albert Bandura. [n. d.]. Self-efficacy: Toward a Unifying Theory of Behavioral Change. ([n. d.]).

[4] Dipto Barman and Owen Conlan. 2024. Evaluating Prebunking and Nudge Techniques in Tackling Misinformation: A Between-Subject Study on Social Media Platforms. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. ACM, Poznan Poland, 167–177. doi:10.1145/3648188.3675125

[5] Melisa Basol, Jon Roozenbeek, and Sander van der Linden. 2020. Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition* 3, 1 (Jan. 2020), 2. doi:10.5334/joc.91

[6] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. 2015. Package 'lme4'. *convergence* 12, 1 (2015), 2.

[7] Jonas Becker. 2024. Multi-Agent Large Language Models for Conversational Task-Solving. arXiv:2410.22932 (Nov. 2024). doi:10.48550/arXiv.2410.22932

[8] Mingwei Chen, Pei-Luen Patrick Rau, and Liang Ma. 2025. Experience Through Flipped Interaction. *Human-Computer Interaction & Emerging Technologies* 195 (2025), 285.

[9] Yang Cheng. 2025. Navigating vaccine misinformation: a study on users' motivations, active communicative action and anti-misinformation behavior via chatbots. *Online Information Review* 49, 3 (2025), 643 – 663. doi:10.1108/OIR-07-2024-0425 Cited by: 0.

[10] Carlos Cinelli, Andrew Forney, and Judea Pearl. 2024. A crash course in good and bad controls. *Sociological Methods & Research* 53, 3 (2024), 1071–1104.

[11] Jacob Cohen. 2009. *Statistical power analysis for the behavioral sciences* (2. ed., reprint ed.). Psychology Press, New York, NY.

[12] Josh Compton. 2024. Inoculation theory. *Review of Communication* (July 2024), 1–13. doi:10.1080/15358593.2024.2370373

[13] Josh Compton and Bobi Ivanov. 2017. Inoculation messaging. In *Persuasion and communication in sport, exercise, and physical activity*. Routledge, 73–90.

[14] Josh Compton, Ben Jackson, and James A. Dimmock. 2016. Persuading Others to Avoid Persuasion: Inoculation Theory and Resistant Health Attitudes. *Frontiers in Psychology* 7 (Feb. 2016). doi:10.3389/fpsyg.2016.00122

[15] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (3 2023), e0279720. doi:10.1371/journal.pone.0279720

[16] Amir Ebrahimi Fard and Shajeeshan Lingeswaran. 2020. Misinformation Battle Revisited: Counter Strategies from Clinics to Artificial Intelligence. In *Companion Proceedings of the Web Conference 2020*. ACM, Taipei Taiwan, 510–519. doi:10.1145/3366424.3384373

[17] Angelo Fasce, Mirela Mustaţă, Alexandra Deliu, Dawn Holford, Linda Karlsson, Virginia Gould, Gheorghe Gindrovel Dumitra, Dana Farcasanu, Iulia Vişinescu, Pierre Verger, and Stephan Lewandowsky. 2025. A field test of empathetic refutational and motivational interviewing to address vaccine hesitancy among patients. *npj Vaccines* 10, 1 (2025), 142. doi:10.1038/s41541-025-01197-8

[18] Maria Mercedes Ferreira Caceres, Juan Pablo Sosa, Jannel A Lawrence, Cristina Sestacovschi, Atiyah Tidd-Johnson, Muhammad Haseeb Ui Rasool, Vinay Kumar Gadamidi, Saleha Ozair, Krunal Pandav, Claudia Cuevas-Lou, Matthew Parrish, Ivan Rodriguez, Javier Perez Fernandez, South Miami Florida USA Division of Research amp; Academic Affairs, Larkin Community Hospital, Coolidge Antigua Department of Medicine, American University of Antigua, Hialeah Florida USA Family Medicine, Larkin Community Hospital Palm Springs Campus, Miami Florida USA Family Medicine, Larkin Community Hospital South Campus, Pulmonary Disease, and Hilaeah Florida USA Critical Care Medicine, Larkin Community Hospital Palm Springs Campus. 2022. The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health* 9, 2 (2022), 262–277. doi:10.3934/publichealth.2022018

[19] Marieke L. Fransen, Saar Mollen, Stephan A. Rains, Enny Das, and Ivar Vermeulen. 2024. Sixty Years Later: A Replication Study of McGuire's First Inoculation Experiment. *Journal of Media Psychology* 36, 1 (Jan. 2024), 69–78. doi:10.1027/1864-1105/a000396

[20] Shixian Geng, Remi Inayoshi, Chi-Lan Yang, Zefan Sramek, Yuya Umeda, Chiaki Kasahara, Arissa J Sato, Simo Hosio, and Koji Yatani. 2025. Beyond the Dialogue: Multi-chatbot Group Motivational Interviewing for Premenstrual Syndrome (PMS) Management. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[21] Louie Giray. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering* 51, 12 (Dec. 2023), 2629–2633. doi:10.1007/s10439-023-03272-4

[22] Linda C Godbold and Michael Pfau. 2000. Conferring resistance to peer pressure among adolescents: Using inoculation theory to discourage alcohol use. *Communication research* 27, 4 (2000), 411–437.

[23] Ziyang Gong and Leona Yi-Fan Su. 2025. Exploring the Influence of Interactive and Empathetic Chatbots on Health Misinformation Correction and Vaccination Intentions. *Science Communication* 47, 2 (2025), 276 – 308. doi:10.1177/10755470241280986 Cited by: 2.

[24] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. 15, 9 (2005), 1277–1288. doi:10.1177/1049732305276687

[25] Youngseung Jeon, Bogoan Kim, Aiping Xiong, Dongwon Lee, and Kyungsik Han. 2021. ChamberBreaker: Mitigating the Echo Chamber Effect and Supporting Information Hygiene through a Gamified Inoculation System. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–26. doi:10.1145/3479859

[26] Loukas Konstantinou and Evangelos Karapanos. 2025. Behavior Change Interventions Combating Online Misinformation: A Scoping Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3713127

[27] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[28] Miikka Kuutila, Carita Kiili, Reijo Kupiainen, Eetu Huusko, Junhao Li, Simo Hosio, Mika Mäntylä, Julie Coiro, and Kristian Kiili. 2024. Revealing complexities when adult readers engage in the credibility evaluation of social media posts. *Computers in Human Behavior* 151 (Feb. 2024), 108017. doi:10.1016/j.chb.2023.108017

[29] Tuan-He Lee and Susan R. Fussell. 2025. Countering Misinformation in Private Messaging Groups: Insights From a Fact-checking Chatbot. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP10 (Jan. 2025), 30 pages. doi:10.1145/3701189

[30] Junhao Li, Eetu Huusko, Nazanin Nakhaie Ahooie, Miikka Kuutila, Aleksi Huotala, Mika Mäntylä, and Simo Hosio. 2025. Credtwi: Investigating Social Media Credibility with a Browser Plugin. *International Journal of Human–Computer Interaction* (April 2025), 1–17. doi:10.1080/10447318.2025.2480885

[31] Samira Malek, Christopher Griffin, Robert Fraleigh, Vishal Monga, and Lijiang Shen. [n. d.]. A Methodology Framework for Analyzing Health Misinformation to Develop Inoculation Intervention Using Large Language Models: A case study on covid-19. ([n. d.]).

[32] Alicia M Mason and Claude H Miller. 2016. Potentially deceptive health nutrition–related advertising claims: The role of inoculation in conferring resistance. *Health Education Journal* 75, 2 (2016), 144–157. arXiv:https://doi.org/10.1177/0017896915569365 doi:10.1177/0017896915569365

[33] Jonathan Matusitz and Gerald-Mark Breen. 2010. Inoculation Theory: A Framework for the Reduction of Skin Cancer. *Journal of Evidence-Based Social Work* 7, 3 (June 2010), 219–234. doi:10.1080/19371910902911172

[34] William J. McGuire. 1961. The Effectiveness of Supportive and Refutational Defenses in Immunizing and Restoring Beliefs Against Persuasion. *Sociometry* 24, 2 (June 1961), 184. doi:10.2307/2786067

[35] William J. McGuire. 1964. *Some Contemporary Approaches*. Vol. 1. Elsevier, 191–229. doi:10.1016/S0065-2601(08)60052-0

[36] Aly Passanante, Ed Pertwee, Leesa Lin, Kristi Yoonsup Lee, Joseph T Wu, and Heidi J Larson. 2023. Conversational AI and Vaccine Communication: Systematic Review of the Evidence. *Journal of Medical Internet Research* 25 (Oct. 2023), e42758. doi:10.2196/42758

[37] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (5 2017), 153–163. doi:10.1016/j.jesp.2017.01.006

[38] Wei Peng, Hee Rin Lee, and Sue Lim. 2024. Leveraging Chatbots to Combat Health Misinformation for Older Adults: Participatory Design Study. *JMIR Formative Research* 8 (Oct. 2024), e60712. doi:10.2196/60712

[39] James W. Pennebaker, Ryan L. Boyd, Roger J. Booth, Ashwini Ashokkumar, and M. E. Francis. 2022. *Linguistic Inquiry and Word Count: LIWC-22*. https://www.liwc.app

[40] Michael Pfau, Steve Van Bockern, and Jong Geun Kang. 1992. Use of inoculation to promote resistance to smoking initiation among adolescents. *Communications Monographs* 59, 3 (1992), 213–230.

[41] Michael Pfau and Michael Burgoon. 1988. INOCULATION IN POLITICAL CAMPAIGN COMMUNICATION. *Human Communication Research* 15 (1988), 91–111. https://api.semanticscholar.org/CorpusID:143574957

[42] Johnmarshall Reeve. 2009. Why Teachers Adopt a Controlling Motivating Style Toward Students and How They Can Become More Autonomy Supportive. *Educational Psychologist* 44, 3 (July 2009), 159–175. doi:10.1080/00461520903028990

[43] Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (June 2019), 65. doi:10.1057/s41599-019-0279-9

[44] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. 2020. Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review* (Feb. 2020). doi:10.37016//mr-2020-008

[45] Alexander Rossner, Marie Ambach, and Sven Pagel. 2025. Navigating Misinformation: Understanding User Frustration in AI-Driven Chatbot Interactions. *Lecture Notes in Computer Science* 15820 LNAI (2025), 263 – 281. doi:10.1007/978-3-031-93415-5_16 Cited by: 0.

[46] Richard M Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology* 43, 3 (1982), 450.

[47] Briony Swire-Thompson, David Lazer, et al. 2020. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health* 41, 1 (2020), 433–451.

[48] Huiyun Tang and Anastasia Sergeeva. 2025. Shots and Boosters: Exploring the Use of Combined Prebunking Interventions to Raise Critical Thinking and Create Long-Term Protection Against Misinformation. arXiv:2505.07486 (May 2025). doi:10.48550/arXiv.2505.07486

[49] Huiyun Tang, Songqi Sun, Kexin Nie, Ang Li, Anastasia Sergeeva, and Ray LC. 2025. Breaking the News: A LLM-based Game where Players Act as Influencer or Debunker for Raising Awareness About Misinformation. arXiv:2502.04931 (Feb. 2025). doi:10.48550/arXiv.2502.04931

[50] John Wilder Tukey et al. 1977. *Exploratory data analysis*. Vol. 2. Springer.

[51] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–35.

[52] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382 (Feb. 2023). doi:10.48550/arXiv.2302.11382

[53] Bingyu Yan, Xiaoming Zhang, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, and Chaozhuo Li. 2025. Beyond Self-Talk: A Communication-Centric Survey of LLM-Based Multi-Agent Systems. arXiv:2502.14321 (Feb. 2025). doi:10.48550/arXiv.2502.14321

[54] Rui Ye, Xiangrui Liu, Qimin Wu, Xianghe Pang, Zhenfei Yin, Lei Bai, and Siheng Chen. 2025. X-MAS: Towards Building Multi-Agent Systems with Heterogeneous LLMs. arXiv:2505.16997 (May 2025). doi:10.48550/arXiv.2505.16997

[55] Jennifer Zarzosa and Cecilia Ruvalcaba. 2025. Fighting Fake News: Building Cognitive Resistance and Mental Antibodies Through a Digital Media Literacy Inoculation Intervention. *Journal of Marketing Education* (Feb. 2025). doi:10.1177/02734753251319559

[56] Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward Mitigating Misinformation and Social Media Manipulation in LLM Era. In *Companion Proceedings of the ACM Web Conference 2024*. ACM, Singapore Singapore, 1302–1305. doi:10.1145/3589335.3641256

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. arXiv:2309.11998 (March 2024). doi:10.48550/arXiv.2309.11998