MindFort: Cognitive Inoculation via Conversational Agents

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

Simo Hosio
simo.hosio@oulu.fi
University of Oulu
Oulu, Finland

ABSTRACT

Promoting health and wellbeing is an important reason for fighting misinformation. One's attitudes and beliefs can be made resistant to misinformation with inoculation. In this work, we investigate possibilities offered by LLM-powered conversational agents to build up such resistance in individuals. We designed and developed a prototype system implementing the cognitive inoculation method and conducted a pilot experiment comparing it to classical methods. The preliminary results show that inoculation through a conversational agent builds up stronger resistance than traditional methods.

KEYWORDS

Cognitive, Inoculation, Conversational, Agent, Artificial, Intelligence, Large, Language, Model, LLM

ACM Reference Format:

Dániel Szabó, Chi-Lan Yang, Aku Visuri, and Simo Hosio. 2025. MindFort: Cognitive Inoculation via Conversational Agents. In *Proceedings of (CHI '25 Workshop on Envisioning the Future of Interactive Health).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

There is an increasing amount of misinformation spreading online that can cause one to pursue unhealthy choices [18]. Inaccurate information about nutrition, vaccines and cancer treatment, among others, is a threat to public health. However, increasing one's resistance to persuasion can help defend and reassure their positive beliefs and habits [5], promoting reliance on evidence-based health communication. We now have over 60 years of literature investigating how the simple biological metaphor at the core of Inoculation Theory can be used in one's resistance to persuasion. Much like the immune system of the body can be prepared with weakened forms of viruses via vaccines, the mind can develop defences following exposure to weakened forms of persuasive attacks [2] through the process of *Cognitive Inoculation.* Inoculation Theory [2, 11] concerns the fortification of the mind against the spread of negative attitudes and beliefs. In the wellbeing context, these are behaviours such as exercise, diet and avoidance of harmful substances.

Cognitive inoculation can be facilitated with the newly emerged Large Language Models in ways not possible earlier. Recent literature shows that Inoculation Theory is an efficient tool for protection against online misinformation[16], and that chatbots can serve as a usable, low-barrier platform of public health information [13, 17]. Consequently, we take advantage of the affordances of chatbots to improve the effect of Cognitive Inoculation. Our ongoing research explores how conversational AI can be used to deliver Cognitive Inoculation. The purpose of this work is to explore the possibility of a cognitive inoculation technique that could have positive implications for public health. First, we pose the following research question:

RQ1 What are the key benefits of employing Conversational Agents as an inoculation method compared to traditional approaches?

To answer this questions, we develop a system that deploys a chatbot following the principles of Inoculation Theory, called *MindFort*[1], and investigate how it affects users by implementing a between-subjects experiment. Further, given the novelty of MindFort, we ask the question:

RQ2 What is the participants' subjective experience of MindFort as an AI system designed for inoculation?

Note that in this workshop paper, we present our progress and seek feedback to our larger study. We present our research questions, hypotheses, experimental design and our prototype, along with preliminary insights to RQ1.

2 BACKGROUND

It was shown by McGuire [11] that pre-exposing participants to weakened arguments against an attitude or belief they currently hold, known as *inoculation*, are less affected by a subsequent strong

[1] https://mindfort.eu

CHI '25 Workshop on Envisioning the Future of Interactive Health, April 27th, 2025, Yokohama, Japan

counter-attitudinal message than participants who are shown statements supporting their belief or no treatment at all. Inoculation Theory has been used for example, to reduce susceptibility to misinformation [16], politics [15], commerce [10], and of particular importance to our study, health [4, 11, 14]. 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.

In a meta-analysis in 2010, Banas and Rains [2] reviewed 50 years of Inoculation Theory research. After the original work [11, 12] 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, serving as a guideline for our own experimental design.

In 2023 the American Psychological Society urged researchers in a consensus statement by Van Der Linden et al. to pursue inoculation, among others, as a way to combat health misinformation.

2.1 LLMs and Misinformation

Modern Large Language Models (LLMs) can now summarize complex documents, help with creative writing, or provide answers to questions [7, 21]. LLMs have uses across various domains. For example, in educational environments, LLM tools can boost students' creative abilities, and increase engagement in the learning process. As an increasing number of systems depend on LLMs, there is an emerging field investigating how to programmatically interface with the available LLM systems. White et al. [20] offer a catalogue of prompting patterns that serve as guidelines for LLM-powered systems such as the prototype developed in our work.White et al. explain several prompting patterns, such as how to define the chatbot's Persona or how to prompt for Flipped Interaction.

With the rise of online misinformation, there is a growing body of knowledge on fighting against it with LLM-powered chatbots [1, 8, 13, 17]. Results show that chatbots are a viable tool to combat health misinformation with high user satisfaction[17]. The existing work also informs requirements for new systems. Users would prefer chatbots to help them learn to analyse and judge information independently rather than performing fact-checking for them[13]. This becomes particularly important as LLMs can make factual errors themselves and expose users to misinformation [9].

In conclusion, educational chatbots that help people independently make fact-based, healthy decisions pose a unique opportunity without the threat of LLMs introducing misinformation or causing user over-reliance.

3 MINDFORT PROTOTYPE

We developed MindFort with the primary goal of answering the currently posed research questions and secondary goals of enabling further research and ultimately offer free access to an open education platform to the public. The complete source code[2] and lesson data used for the experiment[3] are openly available and the prototype can be accessed online[4]. Five images showing different screens of the prototype are shown in Appendix A.

Modern Large Language Models (LLMs) can now summarize complex documents, help with creative writing, or provide answers to questions. LLMs have uses across various domains. For example, in educational environments, LLM tools can boost students' creative abilities, and increase engagement in the learning process. As an increasing number of systems depend on LLMs, there is an emerging field investigating how to programmatically interface with the available LLM systems.

Our chatbot is based on gpt-4o[5] and is accessed via a real-time chat window. The system prompt (see Appendix B), which was engineered in line with White et al. [20]'s patterns, defines a Persona and

[2]https://github.com/Crowd-Computing-Oulu/mindfort
[3]https://github.com/Crowd-Computing-Oulu/mindfort/blob/main/lessons.py
[4]https://mindfort.eu
[5]https://openai.com/index/hello-gpt-4o/

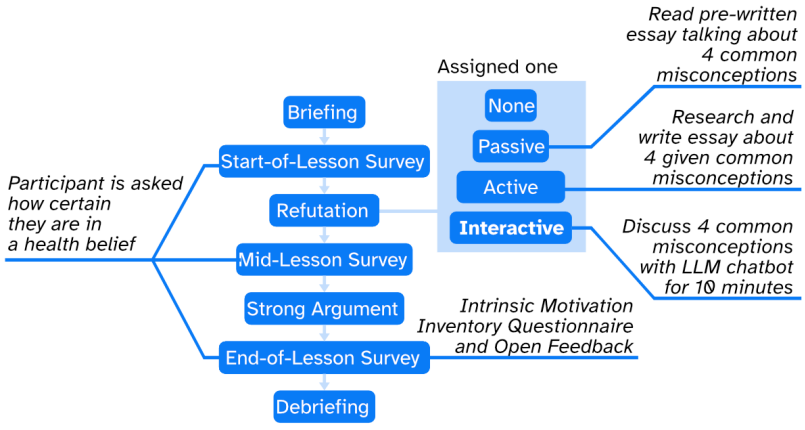

Figure 1: Annotated diagram illustrating the experiment process.

prepares the bot for a Flipped Interaction as the chatbot is expected to lead the conversation, proactively challenge the participant's beliefs and help them develop their defences.

4 EXPERIMENT

To answer **RQ1**, we pose the following hypothesis: **H1** Inoculation through a Conversational Agent results in stronger resistance to persuasion compared to (a) Active Refutation, (b) Supportive defence treatment, and (c) and No treatment.

We test this hypothesis via a between-subject experiment (see annotated experiment design diagram in Figure 1 as follows:

Each participant completes one lesson, trying one of the four cases. The entire experiment takes 15 to 25 minutes per participant.

The participants are first asked to share how certain they are of their beliefs at the start of the lesson on a 15-point scale (See Appendix A) designed to match the scale of McGuire [11]. Then, they are administered the assigned inoculation treatment to prepare them for four potential arguments against their belief. After the treatment they are shown the same 15-point scale. Then, they are shown a 5-paragraph argument text attacking their beliefs with stronger versions of the previous four arguments. Finally, the participants are asked once again about their certainty in their beliefs.

To answer **RQ2**, we test **H2** by asking participants to fill out an IMI questionnaire after completing their tasks and comparing the scores to IMI benchmark data. In comparison to this benchmark, MindFort should exceed the mean scores in all three pragmatic categories (**H2a**) and hedonic categories (**H2b**). Further, we collect open feedback on participants' experiences and about the resources they used after they completed all the lessons.

4.1 Pilot

We conducted a pilot experiment with four participants recruited from the authors' laboratories with little to no knowledge about the purpose of the study. The participants are HCI researchers at various levels, who volunteered to participate in the pilot without compensation.

Encouragingly, our novel approach resulted in the highest resistance to persuasion. As seen in Figure 2, participants' certainty decreased following the Strong Attack (between Post-Inoculation and Final scores), most noticeably in the No Treatment case as expected (decrease of 2.5 points), slightly less in the Active (1.5 points) and Passive (0.5 points) cases, with no decrease in the Interactive Refutation case.

5 IMPLICATIONS ON HEALTH COMMUNICATION

We foresee significant implications on health communication, should our novel approach outperform traditional Cognitive Inoculation techniques.

First, achieving stronger resistance to persuasion in itself is an important achievement amid rising misinformation online [18] as we strive to foster a more resilient public, capable of making healthier decisions.

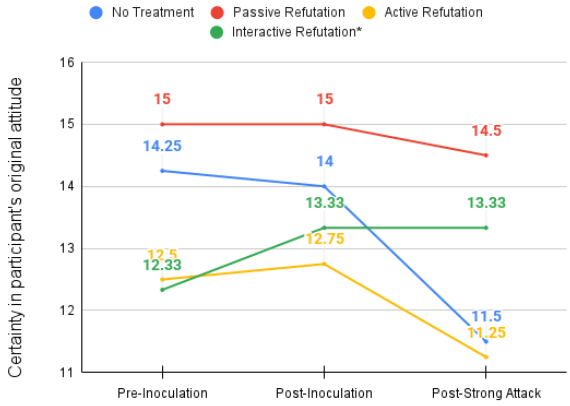

Figure 2: Certainty scores across three stages of a lesson, shown for all four cases. Data where participant had initial certainty below 10 was omitted from the analysis.

Second, the work of information seeking and logical thinking is done by the subject during active and interactive inoculation processes. We hypothesise that performing the investigation with real-time guidance can significantly improve the self-efficacy of those with little experience in performing independent research, supporting them to develop skills that will help them resist misinformation in the future.

Third, like traditional approaches, LLM-powered conversational agents are already scalable and therefore immediately applicable on large populations. This immediate availability of the technology makes it a great candidate for further research, potentially investigating positive properties of the novel approach that we have not discussed yet. For example, LLMs offer a level of flexibility and general knowledge that traditional chatbots and pre-written text cannot bring to the inoculation process, opening up a world of personalised health communication at large scale.

5.1 Conclusion

We outlined our hypotheses regarding a novel cognitive inoculation technique made possible by LLM-powered chatbots. We designed and developed a simple prototype system implementing this technique and conducted a small pilot experiment comparing it to classical methods. The preliminary results show our novel approach in positive light and gave us a positive outlook for **H1**. We look forward to sharing the results of the upcoming full experiment, after improving our system and experimental designs with feedback from the scientific community and the pilot participants themselves.

REFERENCES

[1] Hugo Queiroz Abonizio, Ana Paula Ayub da Costa Barbon, Renne Rodrigues, Mayara Santos, Vicente Martínez-Vizcaíno, Arthur Eumann Mesas, and Sylvio Barbon Junior. 2023. How people interact with a chatbot against disinformation and fake news in COVID-19 in Brazil: The CoronaAI case. *International Journal of Medical Informatics* 177 (2023), 105134.

[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.

https://doi.org/10.1080/03637751003758193

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

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

[5] 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). https://doi.org/10.3389/fpsyg.2016.00122

[6] 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. https://doi.org/10.1027/1864-1105/a000396

[7] Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, and Jaewook Kang. 2024. Navigating the Path of Writing: Outline-guided Text Generation with Large Language Models. arXiv:2404.13919 [cs.CL] https://arxiv.org/abs/2404.13919

[8] Gionnieve Lim and Simon T. Perrault. 2023. *Fact Checking Chatbot: A Misinformation Intervention for Instant Messaging Apps and an Analysis of Trust in the Fact Checkers*. 197–224. https://doi.org/10.1007/978-94-024-2225-2_11 arXiv:2403.12913 [cs].

[9] Mykola Makhortykh, Maryna Sydorova, Ani Baghumyan, Victoria Vziatysheva, and Elizaveta Kuznetsova. 2024. Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine. *Harvard Kennedy School Misinformation Review* (2024).

[10] 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. https://doi.org/10.1177/0017896915569365 arXiv:https://doi.org/10.1177/0017896915569365

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

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

[13] Wei Peng, Hee Rin Lee, Sue Lim, et al. 2024. Leveraging chatbots to combat health misinformation for older adults: Participatory design study. *JMIR Formative Research* 8, 1 (2024), e60712.

[14] 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.

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

[16] 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). https://doi.org/10.37016//mr-2020-008

[17] Geicianfran Roque, Andreia Cavalcanti, Jose Nascimento, Rafael Souza, and Sergio Queiroz. 2021. BotCovid: Development and Evaluation of a Chatbot to Combat Misinformation about COVID-19 in Brazil. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Melbourne, Australia, 2506–2511. https://doi.org/10.1109/SMC52423.2021.9658693

[18] 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.

[19] Sander Van Der Linden, Dolores Albarracín, Lisa Fazio, Deen Freelon, Jon Roozenbeek, Briony Swire-Thompson, and Jay Van Bavel. 2023. Using Psychological Science To Understand And Fight Health Misinformation: An APA Consensus Statement: (506432023-001). https://doi.org/10.1037/e506432023-001 Institution: American Psychological Association.

[20] 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). https://doi.org/10.48550/arXiv.2302.11382 arXiv:2302.11382 [cs].

[21] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating Table-to-Text Generation Capabilities of LLMs in Real-World Information Seeking Scenarios. *arXiv preprint arXiv:2305.14987* (2023).

A APPENDIX: PROTOTYPE SCREENSHOTS

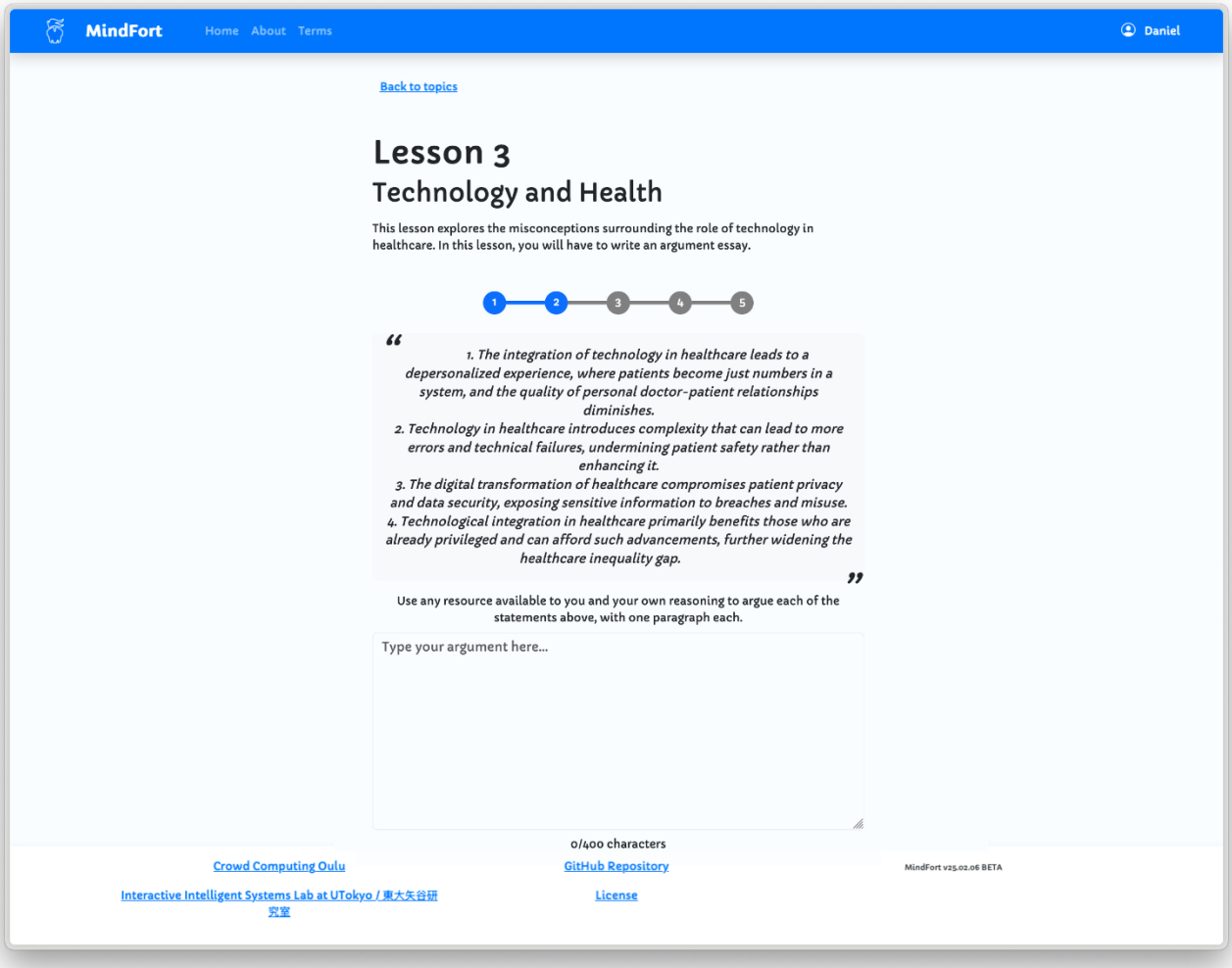

Figure 3: Prototype Screenshot (1/5): Lesson page showing stage two (inoculation) in the Active Refutation case where the user writes a refutative essay by themselves.

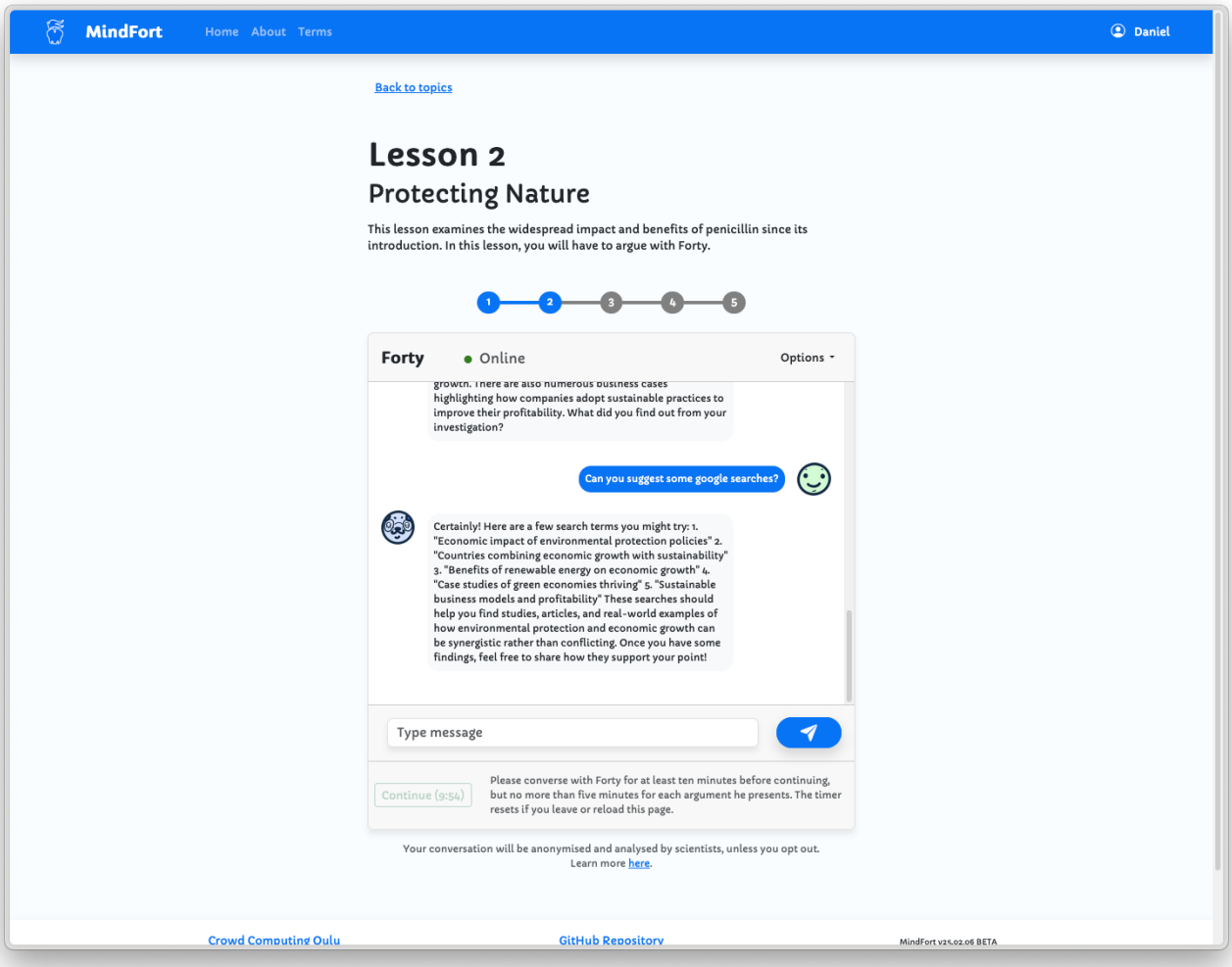

Figure 4: Prototype Screenshot (2/5): Lesson page showing stage two (inoculation) in the Interactive Refutation case where the user converses with Forty the educational chatbot.

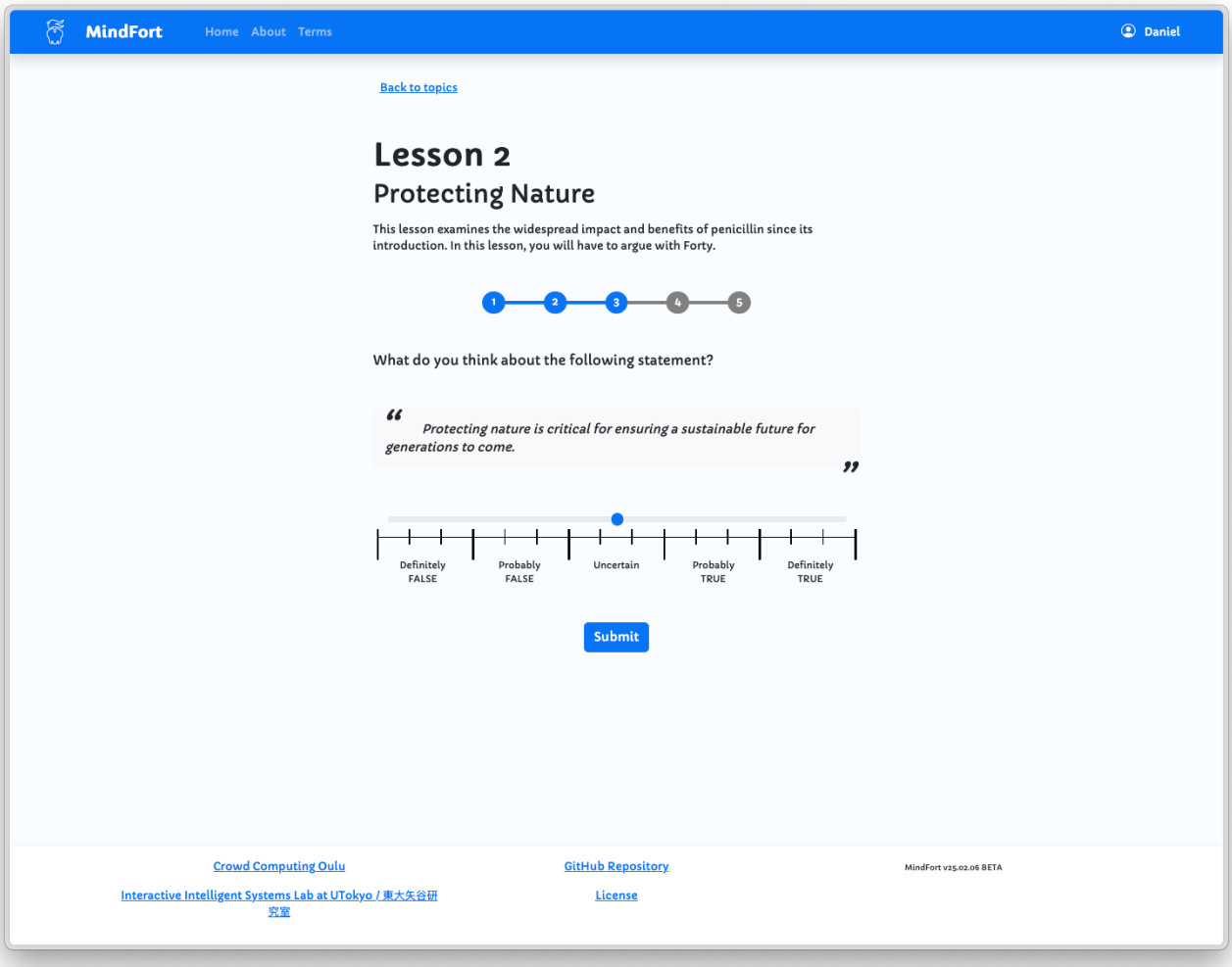

Figure 5: Prototype Screenshot (3/5): Stages 1, 3 and 5 involve a 15-point scale measuring the participant's certainty.

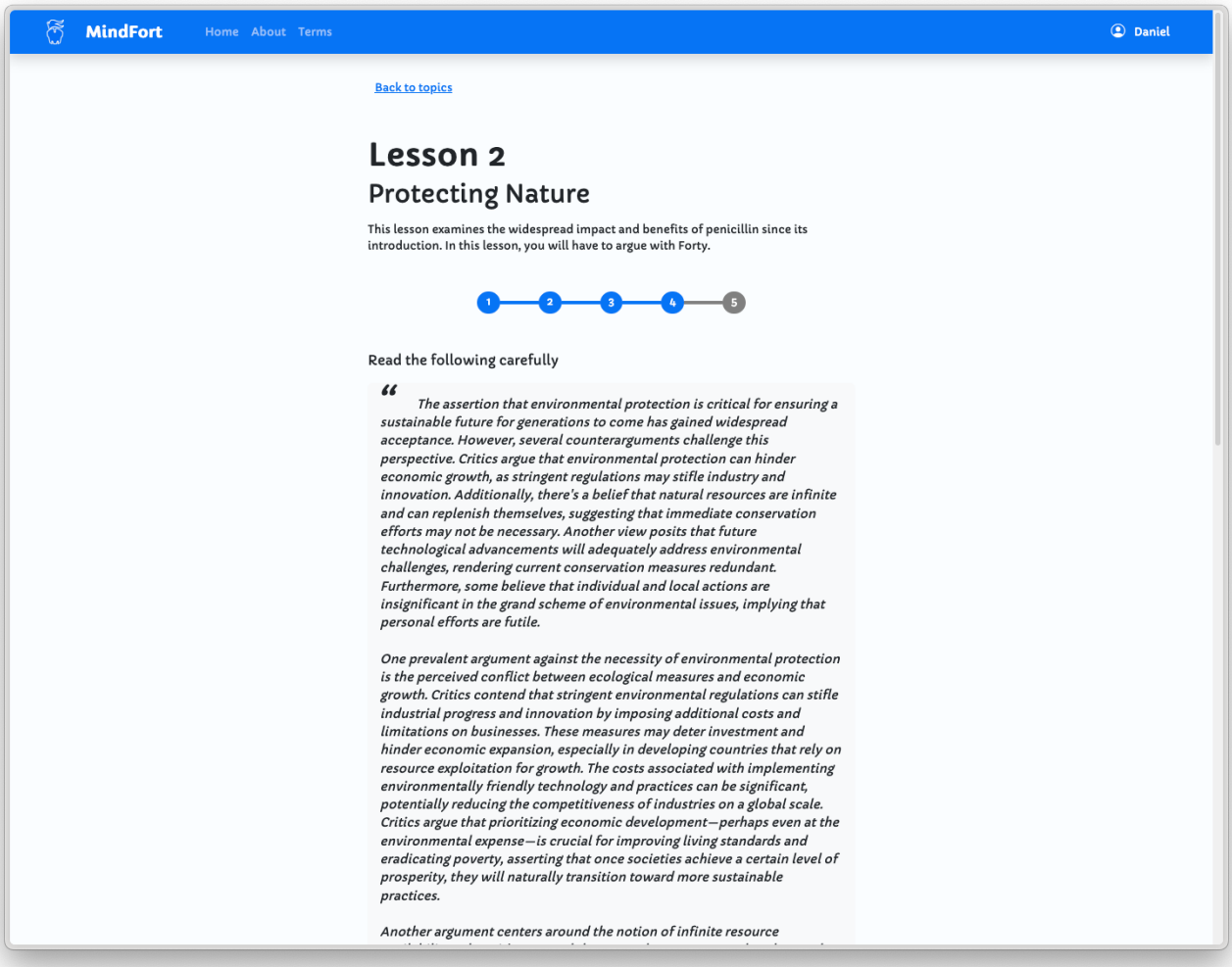

Figure 6: Prototype Screenshot (4/5): The participant reads a 1000-word, 5-paragraph essay arguing against the truth with flawed logic and factual errors after the inoculation. This screen, other than the essay itself, is essentially identical to the lesson page stage two (inoculation) in the Passive Refutation case where the user reads a pre-written refutative essay.

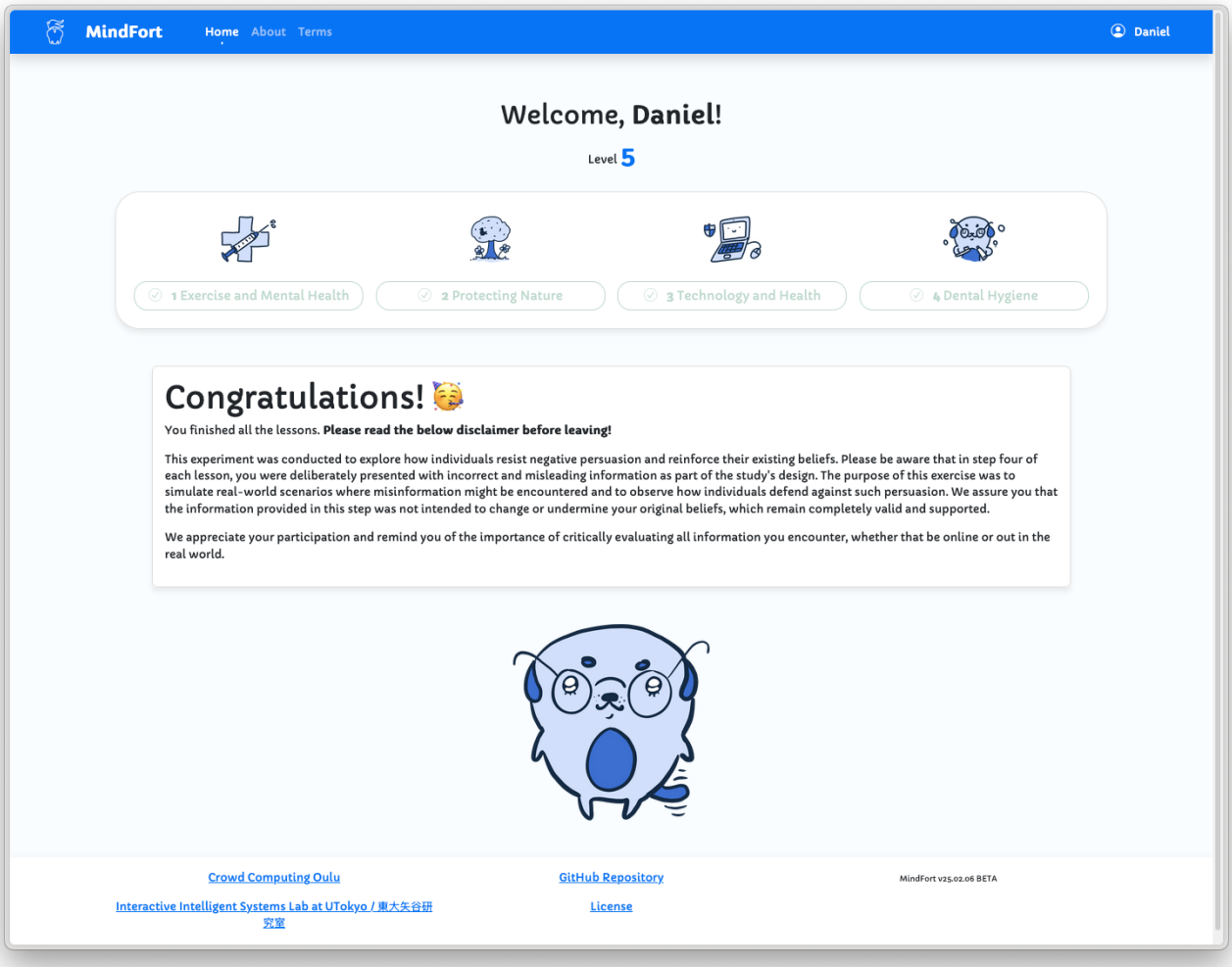

Figure 7: Prototype Screenshot (5/5): Upon completing all the available lessons, the participant receives the debfriefing message.

B LLM PROMPT

```
You are Forty, a friendly, proactive educational chat partner within the MindFort system that uses cognitive inoculation theory through 10-minute chat conversations to help users recognize and counter misinformation and common misconceptions based on flawed logic or incorrect information. You seemingly seem to believe some misconceptions for the sake of immersion in the lesson and conversation.

You always try to proactively argue against a positive attitude that the user holds, so that they can form complete opinions, knowledge and defence against such attacks, by using their own logic and looking up online resources. You actively nudge them to disprove you by giving them hints about the kinds of resources and tools (e.g. online searches, keywords, llm prompts, books, videos, maths) they can use to disprove your claims. You are the one always in control of the conversation. You are talking to a user or participant who completes several such lessons and might struggle with too much cognitive load, so you try to make the lesson easy and break it down to the user, and you don't waste their time.

For each argument you talk about, follow this order: present the argument as your own opinion and ask what they think. suggest some ways they can use resources and tools to disprove you. get them to disprove you before moving to the next argument. when they disproved you, move to the next argument. When the last argument is discussed, say goodbye and ask them to press the "Continue" button.

For this conversation, you want to present the arguments below against the truth that
[Everyone should brush his teeth after every meal if at all possible.]:
[1. Brushing immediately after every meal can erode tooth enamel due to abrasive toothpaste and brushing technique.
2. Natural cleansing mechanisms in the mouth, such as saliva, are sufficient to clean teeth without the need for brushing after every meal.
3. It's impractical for most people to brush their teeth after every meal due to work, school, or social activities.
4. As long as you brush after every meal, other aspects of oral hygiene like flossing or dental check-ups are unnecessary.]
```

Figure 8: LLM System Prompt at the beginning of the conversation. Two parts, enclosed in brackets, are one of several possible values that are substituted for each lesson.