Krillogical: A Collaborative Game On Fallacious Reasoning

Bogdan Badale[1], Kian Sie[1], Sandro Karhula[1], Deniz Çetin[1], Thomas Verwaal[1], David van der Maas[1], Tzu-Hsin Hsieh[1], Amir Zaidi[1], Francesca de Rosa[2,3], and Rafael Bidarra[1]

[1] Delft University of Technology, Delft, Netherlands
R.Bidarra@tudelft.nl
[2] FDR Strategies, Lerici (SP), Italy
[3] Center for Advanced Preparedness and Threat Response Simulation (CAPTRS), Austin (TX), United States

Abstract. Disinformation is becoming pervasive, making it difficult for both individuals and organizations to recognize manipulative or misleading content. Traditional countermeasures such as seminars or courses require a time and resource investment often unaffordable by companies, while text-based materials lack the engagement and interactivity needed for effective learning. To address this gap, we present Krillogical, a serious game designed to train recognition and understanding of informal logical fallacies. The game combines the use and detection of logical fallacies with collaborative play and reflection to build awareness and critical thinking skills. Moreover, the game mentors players using Large Language Model (LLM) feedback on player-created fallacious texts. In a preliminary evaluation, learning outcomes were evaluated with Ricco's Informal Fallacy Task (IFT), comparing fallacy detection skills between players and a control group. We observed average scores of 75% and 76.5% for the test and control groups respectively, indicating that our solution performs on par with conventional teaching methods while illustrating future optimization potential.

Keywords: Serious games · Critical thinking · Disinformation · Large Language Models.

1 Introduction

Digital communication technologies have recently revolutionized the propagation of information. While mostly beneficial, they are also used by malicious actors to spread disinformation: misleading information intended to deceive, manipulate or cause harm. Disinformation is now created at an unprecedented rate by actors to shape public opinions and undermine trust in institutions and often contains persuasive techniques like emotional language and repetition [15]. Social media disinformation during the COVID-19 pandemic, for example, led to global confusion, mistrust and hindered public health efforts [2].

Disinformation affects both external reputation and internal decision-making of organizations. Therefore, disinformation detection training is becoming increasingly important [4]. Current training techniques primarily consisting of non-interactive seminars, but passive training often results in limited engagement and low learning efficacy [7,10]. In contrast, serious games offer an active form of learning that can overcome these limitations, providing more effective learning, and an enhancement of cognitive abilities and skill development across a wide range of domains [19,9,17]. However, up to the authors' knowledge, no existing game specifically trains players in collaborative creation and detection of disinformation. In fact, games commonly involving lying or employing disinformation as a mechanic (e.g. Mafia/Werewolf [3], and Harmony Square [16]) lack reflection and collaboration and furthermore lack a focus on learning.

We introduce *Krillogical*, a serious game designed to improve the ability to recognize disinformation in a training setting. In *Krillogical*, players must both create and identify manipulative strategies in texts, thereby forcing them to critically analyse the mechanics of deception and improving their ability to detect disinformation. Play is structured through three phases: authoring, detection, and recap.

We use a hybrid scoring approach: an initial LLM score with formative feedback, followed by a secondary human intervention step to verify the LLM's evaluation. This hybrid approach reduces reliance on potentially inaccurate LLMs and encourages reflection and collaboration among players, which also aims to balance scoring.

Finally, we describe the initial evaluation of Krillogical's effectiveness at improving fallacy-detection skills compared with traditional training methods. We carried out a user study that measures learning outcomes using the Ricco's Informal Fallacy Task (IFT), which evaluates participants' abilities to identify and explain logical fallacies and contains 6 informal fallacies with 2 arguments per fallacy. We also observed the play sessions and collected feedback to help further improve and evolve the game going forward.

2 Related Work

Although considerable research has been conducted on both misinformation, the unintentional distribution of false information, and disinformation, the intentional distribution of false information, as well as their relationship to critical thinking, the role that games can play in identifying and improving these elements remains unexplored. We focus primarily on disinformation, as *Krillogical* does not facilitate the detection of misinformation, purely disinformation. Despite this, we include several references regarding misinformation, as there is substantial overlap in the critical thinking skills required to evaluate false texts, whether intentionally falsified or not.

We explore existing research about critical thinking and learning interventions in section 2.1 along with existing inoculation and disinformation games in section 2.2.

2.1 Critical Thinking and Intervention

Machete & Turpin [7] conducted a literature review on the role of critical thinking in identifying fake news and found that subjects struggle to distinguish between true and fake news. They also found that both critical thinking and information literacy are essential for evaluating the credibility of online information. Similarly, Redaelli et al. [10] found a strong correlation between the mastery of critical thinking concepts and the ability to identify misinformation and fakeness. However, their intervention, which used non-interactive videos to teach these skills, proved ineffective, possibly because of the passive nature of the videos. Together, these findings emphasize that critical thinking and information literacy are essential tools for resilience against disinformation, but also that more active approaches may be needed to effectively develop such skills.

Abrami et al. [1] conducted a meta-analysis that shows that critical thinking skills can be taught at all educational levels. They identified two types of instructional intervention that were key to improving critical thinking skills: dialogue and exposure to real-world problems. Dialogue was particularly effective with teacher-posed questions and teacher-led class and group discussions. Real-world problems were especially useful with problem-solving and role-playing methods. Overall, the study found that the combination of dialogue, authentic instruction, and mentoring was the most effective at developing critical thinking skills. This suggests that providing group discussion questions for the players at the end of a game, to reflect on their performance, could improve the players' critical thinking skills. Furthermore, dialogue could be incorporated with team-based discussions, and mentoring could be provided by an LLM.

Hruschka and Appel found that learning about informal fallacies through learning intervention improved participants' ability to identify fake news [5]. This indicates that it could be sufficient for a game to develop the players' skills in identifying fallacies to improve their ability to identify fake news. There are two types of fallacies: formal fallacies and informal fallacies. Formal fallacies are errors in reasoning that arise from the structure of an argument; a conclusion is drawn that does not necessarily follow from the premises. Informal fallacies are errors in the reasoning of an argument that follow from its content, context, or ambiguity, rather than a flaw in its logical structure. In such cases, the argument fails to provide enough support for its conclusion.

2.2 Inoculation and Disinformation Games

A promising direction for more active interventions comes from *inoculation theory* [8], which builds resistance to misinformation by exposing individuals to weakened forms of manipulative arguments or tactics. Roozenbeek and van der Linden [13] introduced the "Bad News" game where, by simulating the role of a disinformation agent, players learn about six common disinformation strategies: impersonation, emotional appeals, conspiracy framing, trolling, and group polarization. A large-scale evaluation with over 15,000 players found that such active inoculation improved recognition of manipulation techniques and reduced the perceived credibility of false content across demographic groups. Roozenbeek

and van der Linden [12] also developed a fake news game in which participants create a fake news article about a politicised topic. They found that the game helped improve the participants' critical thinking skills, suggesting that writing disinformation texts is an effective teaching method that we could incorporate in our game. These findings suggest that an inoculation-based approach could be a promising, scalable and proactive way of building misinformation resilience.

Disinformation creation games have also been prevalent in the entertainment gaming industry. In the game "Harmony Square"; adapted from the aforementioned "Bad News" game, players are tasked with sabotaging a local election by spreading discord and fake news [16]. Rozenbeek et al. [14] showed that the game improved players' ability to recognise manipulative posts as less reliable and discouraged them to share such posts. This provides a strong indication that a game-based approach can be very beneficial in fostering peoples' ability to detect unreliable texts. Our approach expands on this idea by incorporating collaborative elements, which were lacking in the previous games. The combination of fallacy creation and detection with teamwork and LLM-supported mentoring allows players to train critical thinking skills through social dialogue and guided reflection.

Building on the related work, we thus propose a serious game that incorporates strongly emphasized collaborative elements to provide an active method to develop the critical thinking required to evaluate the trustworthiness and credibility of a text.

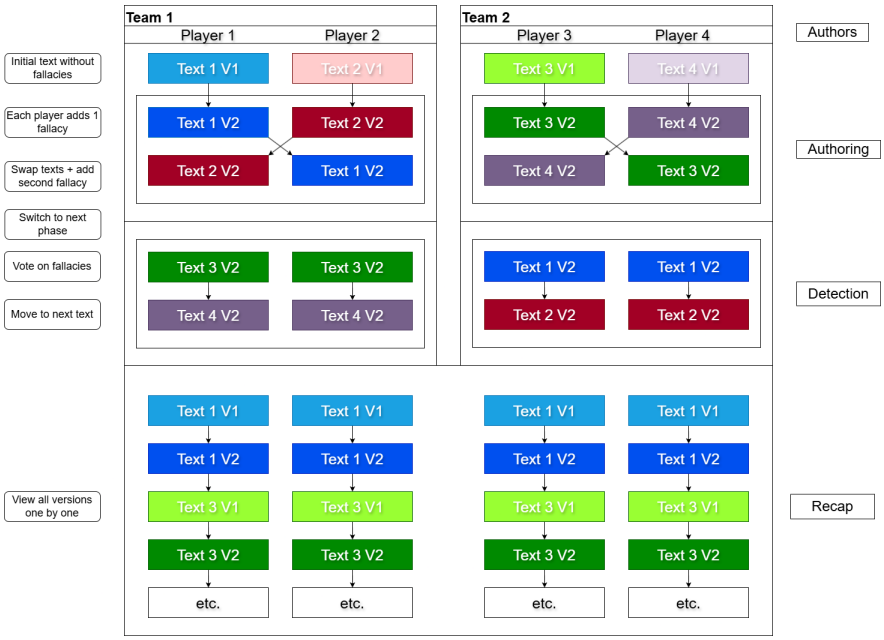

Fig. 1. Flow of a 2-vs-2 game: players add one fallacy per text during *Authoring*; teams identify two fallacies per text during *Detection*; *Recap* guides improvement.

3 Game Design

Krillogical is a team-based serious game designed to improve players' critical thinking skills by letting them engage in the creation, identification, and evaluation of disinformation [7,10,1,5,13,12]. The game is played in two teams with at most three players per team recommended to ensure a smooth and manageable game loop, though larger teams are possible. It contains three phases: Authoring, Detection and Recap (fig. 1). Team members first collaborate to craft misleading texts containing logical fallacies using disinformation strategies. Teams then swap texts and identify fallacies introduced by opponents. Finally, all players review the evolution of the texts, the incorporation of fallacies, the LLM evaluation, and the final scores.

3.1 Authoring Phase

Fallacy *authoring* improves disinformation resilience by enhancing awareness of manipulative reasoning [12]. Players receive a text prompt and choose a fallacy to incorporate. The text is then passed to a teammate who adds another fallacy. This continues until each team member has contributed to all texts belonging to that team, resulting in a set of texts infused with multiple fallacies.

Prompts are general, neutral text prompts to avoid bias and to ensure a focus on reasoning rather than subject knowledge. Two sentences is an optimal prompt length as a single sentence makes it difficult to insert a fallacy, while longer passages make detecting fallacies overly challenging.

Six fallacies were selected, with large (but not complete) overlap with the IFT fallacies, with definitions and examples:

- Appeal to Popularity
- False Cause
- Slippery Slope
- Argument from Ignorance
- Circular Reasoning
- Ad Hominem

Players can view teammates' screens and assist each other which, as shown by Abrami et al. [1], enhances critical thinking skills by stimulating discussion between players. The host, ideally a fallacy expert, can virtually view all screens to monitor progress and provide guidance. A chat is available to support remote discussion, though in-person interaction is preferred. A configurable timer adds pressure, simulating scenarios where experts make decisions in a time-critical environment.

3.2 Detection Phase

The resulting texts from *Authoring* are passed to the opposing team for *Detection*, in which players identify the fallacies added by opponents, to improve the players' ability to identify disinformation, which, as shown by Redaelie et al. [10], has a high correlation with critical thinking skills. For each text, players vote on the fallacy types they believe the opposing team added. Instead of requiring teams to vote collectively, players cast individual votes, allowing within-team disagreements to be settled by distributing votes.

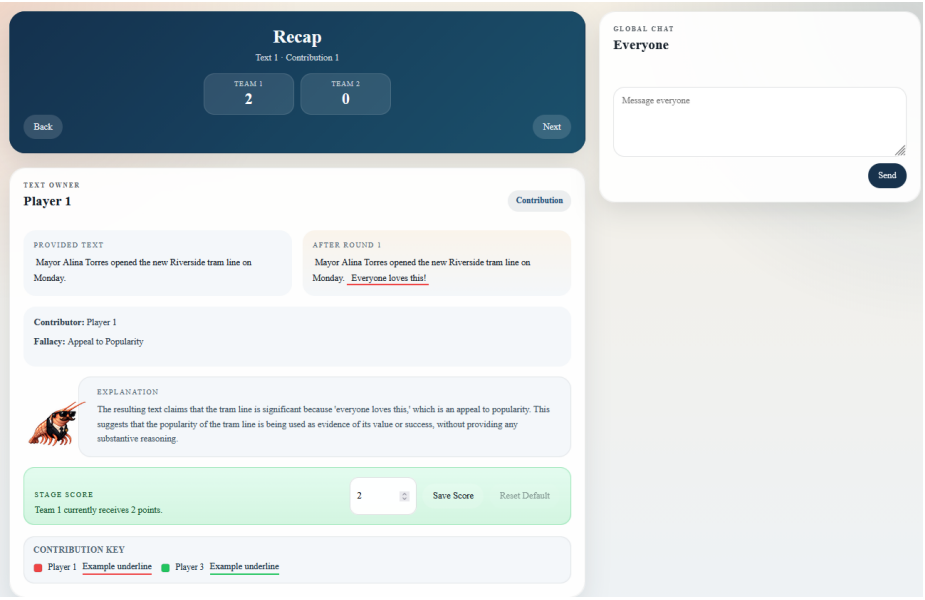

Fig. 2. Recap phase walk-through example: contains the received text, the modifications, the fallacy selected, and the analysis and scoring provided by the LLM. The host ability to override points is also visible as is the global chat players have access to.

3.3 Recap Phase

Finally, the game enters *Recap*, where the evolution of all texts is reviewed in sequence by a host, who can also be a player. Text highlighting clearly shows changes, and votes cast by opposing teams are simultaneously displayed. This detailed walk-through allows players to see how fallacies were incorporated and how peers responded, encouraging reflection on both correct and incorrect identifications. Players gain a concrete understanding of how disinformation evolves through edits and peer evaluation, while the host guides discussion to clarify misunderstandings and highlight key strategies, similar to teacher-led classroom discussions. By facilitating discussion and reflection, *Recap* reinforces analytical reasoning and critical thinking skills [1,12,10]. The host and the players then proceed to the Summary page, which provides an overview of the final results, including the total score and the winning team. Teams receive points based on the accuracy of their fallacy implementation and on their detection performance:

- Teams receive $1 \times (teamsize)$ **authoring points** for each "correctly" applied fallacy.
- Teams receive **1 detection point** for each "correctly" identified fallacy by a player, with full points awarded for correctly identifying all implemented fallacies even if there are less than they should be or if none were implemented (empty submission).

This method of scoring places emphasis on collaboration and prevents players from not submitting texts to withhold points from the opposing team. The

ground truth of the existence of a fallacy in a text, used to determine how many points team should receive, is determined by the LLM. This is, however, not foolproof. To address this limitation, the host can override the LLM's scoring when players point out mistakes in its evaluation. Beyond ensuring fairness and balance, this mechanism also stimulates discussion between players.

4 Evaluation

To evaluate the efficacy of *Krillogical* at teaching players about logical fallacies, we utilise Ricco's Informal Fallacy Task (IFT) [11]. Participants are presented with each of the arguments in a random order and is tasked with identifying whether a fallacy is present, and if so, explaining what fallacy is present. We use the IFT's original criteria for awarding half a point or a full point for each answer. Hruschka et al. [5] utilised the IFT to evaluate the efficacy of their critical thinking learning intervention. They found that the ability to identify fallacious arguments improved the participant's ability to identify fake news, which suggests that this could also apply to *Krillogical*. Our group of play testers consists primarily of university students aged 20-25. We divide play testers into two independent groups. No personal identifying information is stored at any point in the testing. The experimental group plays our game for 30 minutes before taking the IFT, while the control group is given 30 minutes to study traditional teaching materials about fallacies before taking the IFT. The control group has access to a textual overview of logical fallacies; a short instructional video and a fallacy-identification quiz [6]. We evaluate the effectiveness of *Krillogical* in comparison to that of those traditional teaching methods.

5 Results and Discussions

We present our findings from comparing two groups as described in section 4: one group playing *Krillogical*, and the other studying traditional learning materials on logical fallacies. This is a preliminary test with a small test sample, however we believe it still gives a promising indication of the game's usefulness and adequacy.

5.1 Preliminary testing results

Participants' ability to identify and explain fallacies was measured through their scores on the IFT. The experimental group, consisting of 9 players, who played *Krillogical*, achieved an average score of 75.0% ($M = 9.00$, $SD = 1.58$), compared to 76.7% ($M = 9.20$, $SD = 2.84$) in the control group which consisted of 5 people. This 1.7% difference (0.2 points on a 12-point scale) was not statistically significant according to Welch's two-sample t test [18], $t(5.41) = -0.15$, $p = .89$, with a 95% confidence interval for the mean difference of $[-3.66, 3.26]$. The effect size was negligible ($d = -0.09$), indicating that no difference in learning outcomes was detected. Further, the average accuracy was nearly identical between both groups. Qualitative observations suggest that the collaborative environment for both *creating* and *detecting* disinformation increased engagement, while at the same time not sacrificing learning quality, however a larger sample would be needed to say for certain.

5.2 Key Findings, Work in Progress, Limitations

Several noteworthy observations emerged during testing:

- **Key Findings:** We noticed a high level of communication within most groups, bouncing back and forth ideas, especially during detection phase. Players seemed more comfortable with the game once they received the second text, an indicator that repeated play increases competency.
- **Future Evaluation Expansion:** Some play-testers struggled with understanding the mechanics and goal of the game. We addressed this by adding a built-in tutorial and a simplified version of the game, to introduce the necessary concepts and mechanics.
- **Limitations:** The LLM does still sometimes miss or forget fallacies in texts. In the future, we expect, through both further LLM fine-tuning and industry-wide LLM quality increase, that the reliability of the LLM evaluation would also increase significantly.

6 Conclusions

We introduced *Krillogical*, a collaborative serious game to train recognition of disinformation through the creation and analysis of texts containing logical fallacies. Our preliminary study, evaluating *Krillogical*'s educational value and experiential aspects, found equivalent IFT results between *Krillogical* and traditional materials, showing that *Krillogical* can provide a promising gamified alternative to traditional disinformation training materials.

The post-game LLM feedback showcases the potential for new educational avenues. Though human intervention is still required due to the unreliability of current models, as technology advances, so do the learning benefits *Krillogical* can offer. Furthermore, the integration of teamwork and debate was an effective promoter of engagement and sustained cognitive effort. A future evaluation should explore long-term retention effects and cross-cultural applicability.

In summary, *Krillogical* explores the potential of combining gamified learning, collaboration, and AI-supported feedback to cultivate a healthy sensibility towards manipulative uses of logical fallacies.

References

1. Abrami, P.C., Bernard, R.M., Borokhovski, E., Waddington, D.I., Wade, C.A., Persson, T.: Strategies for teaching students to think critically: A meta-analysis. Review of Educational Research **85**(2), 275–314 (2015). https://doi.org/10.3102/0034654314551063
2. Bhattacharya, S., Singh, A.: Unravelling the infodemic: a systematic review of misinformation dynamics during the covid-19 pandemic. Frontiers in Communication **10** (2025). https://doi.org/10.3389/fcomm.2025.1560936
3. Davidoff, D.: Mafia [party game]. Created at Moscow State University (1986), later adapted into Werewolf (1997) by Andrew Plotkin
4. Gardberg, N.A., Colleoni, E., Barnett, M.L., Rehbein, K.: Evaluating organizations in an age of absurdity: How fake news disrupts the relationship between business and society. Business & Society **65**(2), 279–301 (2026). https://doi.org/10.1177/00076503251412639

5. Hruschka, T.M.J., Appel, M.: Learning about informal fallacies and the detection of fake news: An experimental intervention. PLOS ONE **18**(3), 1–14 (03 2023). https://doi.org/10.1371/journal.pone.0283238
6. Learning, L.: Common logical fallacies (2025), `https://courses.lumenlearning.com/englishcomp1/chapter/common-logical-fallacies/`, accessed: 2025-11-06
7. Machete, P., Turpin, M.: The use of critical thinking to identify fake news: A systematic literature review. In: Hattingh, M., Matthee, M., Smuts, H., Pappas, I., Dwivedi, Y.K., Mäntymäki, M. (eds.) Responsible Design, Implementation and Use of Information and Communication Technology. pp. 235–246. Springer International Publishing, Cham (2020)
8. McGuire, W.J.: Inducing resistance to persuasion: Some contemporary approaches. In: Berkowitz, L. (ed.) Advances in Experimental Social Psychology, vol. 1, pp. 191–229. Academic Press (1964). https://doi.org/10.1016/S0065-2601(08)60052-0
9. Nascimento, K.G.d., Ferreira, M.B.G., Felix, M.M.d.S., Nascimento, J.d.S.G., Chavaglia, S.R.R., Barbosa, M.H.: Effectiveness of the serious game for learning in nursing: systematic review. Revista Gaúcha de Enfermagem **42** (2021). https://doi.org/10.1590/1983-1447.2021.20200274
10. Redaelli, S., Biller-Andorno, N., Gloeckler, S., Brown, J., Spitale, G., Germani, F.: Mastering critical thinking skills is strongly associated with the ability to recognize fakeness and misinformation. Frontiers in Education **10**, 1577692 (may 2025). https://doi.org/10.3389/feduc.2025.1577692
11. Ricco, R.B.: Individual differences in the analysis of informal reasoning fallacies. Contemporary Educational Psychology **32**(3), 459–484 (2007). https://doi.org/10.1016/j.cedpsych.2007.01.001, 459
12. Roozenbeek, J., van der Linden, S.: The fake news game: actively inoculating against the risk of misinformation. Journal of Risk Research **22**, 570–580 (02 2018). https://doi.org/10.1080/13669877.2018.1443491
13. Roozenbeek, J., van der Linden, S.: Fake news game confers psychological resistance against online misinformation. Palgrave Communications **5**(1), 65 (2019). https://doi.org/10.1057/s41599-019-0279-9
14. Roozenbeek, J., Maertens, R., McClanahan, W.P., van der Linden, S.: Breaking harmony square: A game that inoculates against political misinformation. Harvard Kennedy School Misinformation Review **1**(8) (2020). https://doi.org/10.37016/mr-2020-47
15. Unger, S., Klapproth, J., Boberg, S., Bösch, M., Vief, N., Stöcker, C., Quandt, T.: Features of disinformation: an expert interview study on the perception of disinformation among political, governmental, media and business elites in germany. Journal of Elections, Public Opinion and Parties **35**(3), 472–494 (2025), https://doi.org/10.1080/17457289.2025.2514199
16. University of Cambridge, U.S. Department of State: Harmony square. `https://harmonysquare.game/` (2020), online disinformation literacy game
17. Vlachopoulos, D., Makri, A.: The effect of games and simulations on higher education: a systematic literature review. International Journal of Educational Technology in Higher Education **14** (07 2017). https://doi.org/10.1186/s41239-017-0062-1
18. Welch, B.L.: The generalization of 'student's' problem when several different population variances are involved. Biometrika **34**(1-2), 28–35 (1947). https://doi.org/10.1093/biomet/34.1-2.28
19. Zhonggen, Y.: A meta-analysis of use of serious games in education over a decade. International Journal of Computer Games Technology **2019**(1), 4797032 (2019). https://doi.org/https://doi.org/10.1155/2019/4797032