Hybrid Intelligence for Logical Fallacy Detection

Mariia Kutepova
University of Groningen
marrrie.23@gmail.com

Khalid Al-Khatib
University of Groningen
khalid.alkhatib@rug.nl

Abstract

This study investigates the impact of Hybrid Intelligence (HI) on improving the detection of logical fallacies, addressing the pressing challenge of misinformation prevalent across communication platforms. Employing a *between-subjects* experimental design, the research compares the performance of two groups: one relying exclusively on human judgment and another supported by an AI assistant. Participants evaluated a series of statements, with the AI-assisted group utilizing a custom ChatGPT-based chatbot that provided real-time hints and clarifications. The findings reveal a significant improvement in fallacy detection with AI support, increasing from an F1-score of 0.76 in the human-only group to 0.90 in the AI-assisted group. Despite this enhancement, both groups struggled to accurately identify non-fallacious statements, highlighting the need to further refine how AI assistance is leveraged.

1 Introduction

The increasing prevalence of misleading information has created an urgent need to improve our ability to detect deceptive content. Faulty reasoning (i.e., fallacies), which drives the spread of misinformation across various discourse domains, poses significant risks to informed decision-making and public discourse (Vrbová et al., 2021; Teneva, 2023). Some researchers even liken this problem to an epidemic (Duarte, 2024). Despite its importance, fallacy detection remains an understudied challenge, with current computational argumentation methods struggling to capture the complexity of deceptive arguments (Goffredo et al., 2022).

Van Eemeren and Verheij (2017) note that fallacies have received limited attention in both formal and computational argumentation research. Recent studies have attempted to bridge this gap by constructing fallacy datasets and developing automatic fallacy identification methods. However, none have explored a Hybrid Intelligence (HI) approach, to the best of our knowledge. HI, which integrates human and AI capabilities, is considered highly promising but requires further empirical research to evaluate its effectiveness across different tasks and domains (Dellermann et al., 2019).

The concept of HI is based on the view that while AI excels in data processing and pattern recognition, it lacks the creativity, empathy, and contextual understanding that humans bring to cognitive tasks. Dellermann et al. (2019) emphasize that HI systems are designed to leverage these complementary strengths, enabling humans and AI to learn from each other and improve over time. This continuous adaptation is particularly crucial for complex tasks such as fallacy detection.

Dellermann et al. (2019) also note that while machine learning and HI are advancing toward real-world applications, the next step is to enhance their problem-solving capabilities. Since HI is still a relatively new concept, further empirical research is necessary to assess its effectiveness across various domains and tasks. Furthermore, although theoretical frameworks for HI exist, practical guidelines for integrating human intuition and creativity with AI's computational power remain unclear. This gap highlights the need for studies that not only develop theoretical insights but also provide empirical validation of the advantages of human-AI collaboration.

This study aims to deepen our understanding of HI in fallacy detection by addressing several key questions: How can existing AI models be effectively integrated into a hybrid intelligent system to assist in this process? How do individuals identify fallacies with AI assistance compared to doing so without AI support? What challenges and limitations arise when employing HI systems for fallacy detection?

By addressing these questions, this research seeks to evaluate the effectiveness of HI in fal-

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 197–208

lacy detection and to provide practical insights into its real-world application. The findings have broader implications for decision-making in education, healthcare, and other fields. The results showed a substantial increase in performance with AI support, rising from an F1-score of 0.76 in the human-only study to 0.90 in the HI study. This highlights the potential of HI to enhance and support human cognitive abilities in complex tasks such as argumentation analysis. All resources developed in this paper are publicly available[1].

2 Related Work

This section reviews recent advancements in computational argumentation and fallacy detection, introduces Hybrid Intelligence (HI) and its potential to enhance decision-making and problem-solving, and discusses the application of HI in improving fallacy detection and argument analysis.

Computational Argumentation and Fallacy Detection Fallacy detection within computational argumentation has gained importance as AI becomes increasingly integrated into daily life and research. Combining AI, linguistics, and logic, computational argumentation analyzes, models, and assesses arguments in natural language; a crucial task in today's information age, where misinformation and faulty reasoning threaten public discourse and decision-making (Sourati et al., 2023). Recent advancements include machine learning models and annotated datasets. For instance, Jin et al. (2022) highlight the limitations of existing models in detecting complex fallacies, while Goffredo et al. (2023) have enhanced fallacy detection in political debates through improved datasets and neural network architectures. Despite these advancements, challenges such as explainability persist, as Sourati et al. (2023) emphasize the need for transparent AI systems to build user trust. Practical applications span education and healthcare, where argumentation systems enhance critical thinking and assist in diagnostic decision-making (Atkinson et al., 2017). Integrating computational argumentation techniques into fallacy detection offers significant progress in understanding and analyzing arguments across various fields.

Hybrid Intelligence Human-AI collaboration has been explored across diverse domains, demonstrating its potential to enhance performance and decision-making. In social chatbots, AI is perceived as a companion providing emotional support (Brandtzaeg et al., 2022), while in mental health, AI enhances empathy in peer-to-peer conversations (Sharma et al., 2023). In creative fields, AI serves as a co-creator, generating new ideas, and techniques such as zero- and few-shot learning show promise despite certain challenges (Dang et al., 2022). In education, AI fosters critical thinking and personalized learning (Markauskaite et al., 2022; Muthmainnah et al., 2022). In customer service, AI improves efficiency by handling routine tasks, allowing human agents to focus on complex issues (Vassilakopoulou et al., 2022). Moreover, Jiang et al. (2022) stress the importance of AI supporting human decision-making without overwhelming users, highlighting clear communication and intuitive design as key factors for successful human-AI collaboration.

HI and Fallacy Detection HI combines human cognitive strengths with AI's computational capabilities to enhance problem-solving and decision-making. Unlike Artificial General Intelligence (AGI), which aims to replicate human cognition, HI focuses on leveraging complementary skills, such as human creativity and empathy, alongside AI's ability to process large datasets. Researchers such as Dellermann et al. (2019) highlight the potential of HI to achieve superior outcomes through collaboration, where humans and AI enhance each other's performance. HI's co-evolutionary nature, where both human and AI agents learn from each other, has proven effective in fields such as digital humanities and education. To our knowledge, no previous work has directly targeted fallacy detection. However, related studies such as Guo et al. (2023) demonstrate how AI chatbots improve students' argumentation skills by providing immediate feedback. Our study builds on these concepts by developing a Hybrid Argumentation Assistant that leverages HI to enhance fallacy detection, combining human intuition with AI capabilities to improve cognitive tasks.

3 Methodology

The methodology of this study is primarily exploratory but incorporates experimental elements, including a between-subjects design inspired by Field and Hole (2023), with two treatment groups: one where participants use an AI assistant (HI) and

[1] https://github.com/marrrie23/Hybrid_Intelligence_Research

one without AI assistance. This design was chosen to effectively assess the impact of AI support on fallacy detection when combined with human reasoning. To evaluate participants' ability to identify logical fallacies, the study employed a range of materials and instruments, including surveys, a custom ChatGPT-based chatbot, and various data analysis techniques. Specifically, two user studies were conducted to collect detailed responses. In the human-only study, participants were asked to identify logical fallacies in given statements, with their performance and reasoning captured through open-ended responses. The HI study followed a similar structure but incorporated interactive elements, allowing participants to engage with a ChatGPT-based chatbot for fallacy detection. This setup enabled the evaluation of AI assistance in supporting reasoning and improving performance. Figure 1 illustrates the overall methodology of this paper.

Task The fallacy detection task involved presenting participants with a statement as input and asking them to determine whether it was logically sound or fallacious.

Data Preparation We utilized the "Logic" dataset introduced by Jin et al. (2022), which contains about 2,300 examples of logical fallacies sourced from educational materials. The dataset includes a diverse range of fallacies, each classified into one of 14 categories, collected from online quiz platforms and websites, with annotations provided by undergraduate students. Table 1 presents the distribution of fallacy categories in the dataset.

To address the uneven distribution of fallacy types, we balanced the dataset by randomly removing some entries to achieve a more even number of arguments per fallacy type. After removing irrelevant columns and filtering arguments based on length, we created a dataset with 1,000 arguments.

Also, we enriched the dataset with high-quality non-fallacious arguments from (Gleize et al., 2019) to ensure having both fallacious and non-fallacious arguments for evaluation. To mitigate potential bias arising from significant differences in argument length, we calculated the maximum (890), minimum (19), and average (131) argument lengths. We then filtered the arguments to include only those within the range of 100 to 160 words. This resulted in a smaller but more homogeneous set of fallacious arguments. The "Miscellaneous" category, which was initially present, was dropped since its arguments fell outside the 100–160 word range. Consequently, the number of fallacy types was reduced from 14 to 13.

Finally, we utilized the filtered arguments and derived 20 subsets, each comprising 10 examples, a mix of fallacious and non-fallacious arguments in a 7:3 ratio. The order of the statements was randomized to prevent bias. This structure ensured diversity and minimized bias, allowing for a more reliable assessment of fallacy detection performance. Table 2 shows the distribution of fallacies in the final dataset used for the user studies, while Table 3 provides an example of one of the sets distributed to participants.

Comparison Elements The primary comparison metric is participants' performance rates in identifying fallacies. Additionally, we examined how requesting AI-generated hints influenced participants' performance. Furthermore, participants were surveyed about their trust in AI and their perception of its transparency.

3.1 Study 1: Human-Only Fallacy Detection

This study investigates how well individuals can detect logical fallacies without AI assistance, serving as a baseline for comparison with AI-supported detection. Participants evaluated a set of arguments, determining the presence of fallacies and providing justifications for their responses. The study aimed to capture human reasoning patterns and establish a reference point for assessing the potential benefits of AI in fallacy detection.

Participants This study focused on fallacy detection using human reasoning alone and involved 60 participants recruited through snowball sampling. The target group consisted of higher education students (bachelor's and master's) and recent graduates in the Netherlands, selected for convenience and ease of recruitment via social media, university mailing lists, and student-oriented platforms such as SurveySwap and SurveyCircle.

Study Design and Implementation A structured study was conducted using a questionnaire administered via the Qualtrics platform [2] to evaluate the effectiveness of human-only fallacy detection. The study included 20 unique sets of arguments, each containing 10 distinct instances. Sixty participants were randomly assigned to review one set, ensuring that each set was evaluated by three participants. For each argument, they answered a binary

[2] www.qualtrics.com

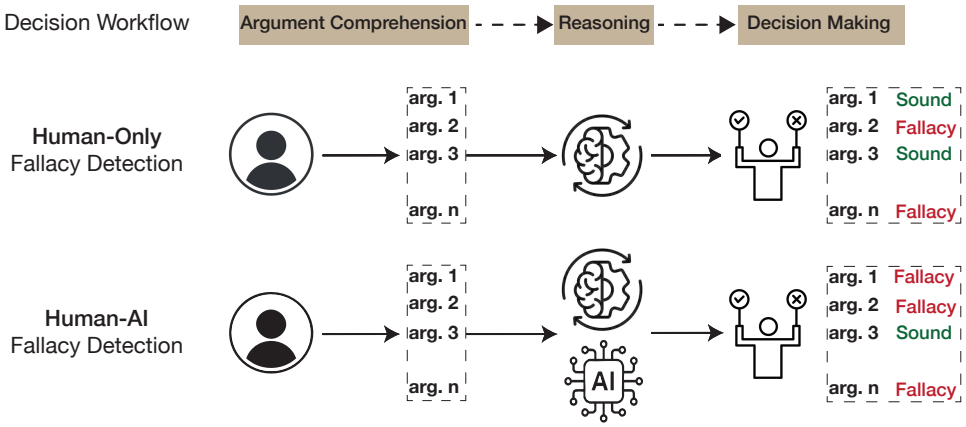

Figure 1: Decision-making workflow for human-only and human–AI fallacy detection.

Fallacy Type	**%**	**Fallacy Type**	**%**
Faulty Generalization	18.4	Intentional Fallacy	5.9
Ad Hominem	12.0	False Dilemma	5.7
Ad Populum	9.3	Fallacy of Credibility	5.7
False Causality	8.6	Fallacy of Extension	5.2
Circular Reasoning	7.1	Equivocation	2.0
Appeal to Emotion	7.2		
Fallacy of Relevance	6.5		
Fallacy of Logic	6.3		

Table 1: Distribution of logical fallacies in (Jin et al., 2022)

question: *"Does this statement contain a fallacy?"* by selecting either "Yes" or "No," followed by an open-ended justification with no constraints. Participants were allowed to complete the questionnaire at their own pace, with no imposed time limits. The data collection period spanned 14 days, providing ample time for participation and ensuring a robust dataset.

3.2 Study 2: Hybrid Intelligence Fallacy Detection

Building on the human-only fallacy detection study, this study explores the potential of AI-assisted reasoning in identifying logical fallacies. By integrating a ChatGPT-based assistant, participants received real-time hints and clarifications to support their decision-making process. This experiment aimed to assess the extent to which AI-generated guidance enhances fallacy detection performance compared to human reasoning alone.

Development of the Hybrid Argumentation Assistant The Hybrid Intelligence (HI) Assistant was developed as a ChatGPT-based chatbot designed to enhance human fallacy detection by providing real-time hints and clarifications. Built using OpenAI's GPT-3.5 model for accessibility, the chatbot interacted with participants by offering targeted hints without explicitly revealing answers, thereby guiding them in identifying logical fallacies.

The chatbot's user interface was designed for seamless interaction, featuring an intuitive chat system that provided feedback and hints. These hints were strategically integrated to encourage deeper reasoning and prompt elaboration on short or insufficient responses, improving both the user experience and the learning process. The chatbot's responses, along with the statements analyzed by participants, were carefully structured within a system prompt to maintain consistency and relevance[3].

Before deployment, the chatbot underwent rigorous testing and refinement to ensure reliability. Initial issues, such as incorrect argument sequencing and missing clarifications, were identified and addressed. The final version was optimized to effectively support participants in accurately detecting logical fallacies, highlighting AI's potential to augment human reasoning.

[3]The system prompt is provided in the appendix

Fallacy Type	Count	Fallacy Type	Count
Faulty Generalization	21	Intentional Fallacy	27
Ad Hominem	24	Fallacy of Credibility	36
Ad Populum	17	False Dilemma	18
False Causality	29	Fallacy of Extension	29
Circular Reasoning	18	Equivocation	23
Appeal to Emotion	27		
Fallacy of Relevance	22		
Fallacy of Logic	23		

Table 2: Fallacy type distribution in the final Dataset

Participants A total of 20 participants were recruited in the same way as in the first study. To ensure high-quality responses, participation was restricted to individuals with a strong performance record on the platform, verified by a high success rate in their previous tasks.

Study Design and Implementation Once the chatbot's behavior aligned with our guidelines, participants were instructed to copy and paste a predefined prompt before starting their interaction with GPT-3.5. After completing their session, they were required to share the link to their full conversation for evaluation.

4 Results

This section presents the findings from the two studies conducted to evaluate the effectiveness of human-only fallacy detection and the impact of AI assistance through a ChatGPT-based HI system. The first study established a baseline for human performance in identifying logical fallacies without AI support, while the second study explored how AI assistance could enhance performance and reasoning in fallacy detection. Both studies utilized the same set of arguments, allowing for a direct comparison of results. The analysis focuses on performance scores, response patterns, and the influence of AI-generated hints, providing insights into the strengths and limitations of human reasoning and the potential of HI in improving fallacy detection.

4.1 Study 1: Human-Only Fallacy Detection

In the first study, 60 participants evaluated logical fallacies without AI assistance. Each was randomly assigned one of 20 argument sets, assessing 10 statements by identifying fallacies (yes/no) and providing explanations. The study analysis followed several key steps to evaluate participants' fallacy detection performance. First, performance was determined based on the effectiveness of fallacy identifications across different argument sets. Responses were compiled into a single dataset for thorough analysis, with performance calculated both per set and per fallacy type. Also, the relationship between short-answer correctness and the length of participants' explanations was explored.

Performance by Set and Fallacy Type The human-only study shows that precision remains consistently high across fallacy types, with most classes reaching perfect precision (1.00). Recall, however, varies more noticeably: high for categories such as *equivocation* (0.92), *appeal to emotion* (0.86), and *fallacy of logic* (0.86), but lower for *false causality* (0.63), *intentional* (0.63), and *false dilemma* (0.64), leading to F1-scores that range from 0.77 to 0.96. Across the 20 evaluation sets, performance is stable, with F1-scores typically falling between 0.70 and 0.84, and only a few sets dipping to 0.62 or rising to 0.93. Overall, these results suggest that humans are highly precise in recognizing fallacies, but their sensitivity varies across types and test sets, highlighting the relative difficulty of consistently detecting certain categories. Table 5 shows fallacy detection performance across the 20 sets, and Table 4 shows the performance across the fallacy types.

Response Length and Performance Correlation The study further examined the relationship between the length of open-ended responses and the performance of fallacy identification. A weak positive correlation (0.325) was observed, suggesting that longer responses were modestly associated with higher performance. While the correlation is weak, it indicates that participants who provided extended responses may have engaged more deeply

Set	Fallacy Type	Argument
S6_A1	no fallacy	"Mandatory vaccinations would limit the spread of the flu and protect vulnerable populations that are at risk of death from the flu."
S6_A2	false dilemma	"Don't waste your money on a home security system; master thieves will still be able to get into your house."
S6_A3	faulty generalization	Donald Trump Jr. Tweeted: If I had a bowl of Skittles and I told you just three would kill you. Would you take a handful? That's our Syrian refugee problem.
S6_A4	no fallacy	Online shopping allows someone in remote areas to access goods normally only available in large cities.
S6_A5	ad populum	I guess I should buy my 12-year-old daughter an iPhone. Everyone at her school has one, and I want her to fit in with the other kids.
S6_A6	equivocation	But professor, I got all these facts from a program I saw on TV once... I don't remember the name of it though.
S6_A7	no fallacy	Addiction to gambling can lead to bankruptcy, families to split, or criminal behavior at times of desperation.
S6_A8	false causality	Mom: Watching TV that close will make you go blind, so move back! Jonny: That is B.S., Mom. Sorry, I am not moving.
S6_A9	circular reasoning	The Senator isn't lying when she says she cares about her constituents—she wouldn't lie to people she cares about.
S6_A10	fallacy of credibility	We should offer movies on our company's website. REPLY: No, we've built our company's fortune by renting movies only through our stores.

Table 3: An example of a set combining high-quality arguments and fallacies.

with the material, leading to higher performance.

Several factors may explain this relationship. First, longer responses might reflect deeper engagement with the arguments, allowing participants to analyze and process the fallacies more thoroughly. Second, participants who wrote more may have had a stronger understanding of the content and the ability to articulate their reasoning more effectively. Lastly, longer responses could indicate greater confidence and familiarity with the topic, enabling participants to provide more comprehensive justifications. In addition, the fallacies "appeal to emotion" and "fallacy of logic", which had the longest average response lengths (391.3 and 373.6 words, respectively), were also associated with higher performance (0.92 and 0.93). This finding supports the idea that more extensive responses may correlate with a deeper understanding or familiarity with the fallacy.

Insights and Implications The variability in performance across sets and fallacy types highlights the complexity of detecting fallacies and suggests gaps in participants' understanding. These findings provide a useful benchmark for evaluating the impact of AI assistance in the next study phase, where participants interact with the HI system. The results also suggest areas where educational interventions may be needed to improve human-only fallacy detection.

4.2 Study 2: ChatGPT-based Hybrid Intelligence Fallacy Detection

In this study, 20 participants used a ChatGPT-based chatbot to assist in fallacy detection. Each analyzed one of the 20 argument sets from the human-only study, identifying fallacies and explaining their reasoning with real-time hints provided by the chatbot. Following data collection, responses were aggregated and analyzed similarly to Study 1. Performance across sets and fallacy types was calculated, and additional variables, such as hint usage, were included to assess the chatbot's impact on fallacy detection. Participants could use hints for both short and open responses, and their usage was tracked for further analysis.

Performance by Set and Fallacy Type The Human–AI study demonstrates consistently strong performance across fallacy categories, with preci-

Fallacy Type	n	Prec.	Rec.	F1
ad hominem	45	1.00	0.82	0.90
ad populum	30	1.00	0.67	0.80
appeal to emotion	21	1.00	0.86	0.92
circular reasoning	33	1.00	0.70	0.82
equivocation	24	1.00	0.92	0.96
fallacy of credibility	36	1.00	0.67	0.80
fallacy of extension	30	1.00	0.67	0.80
fallacy of logic	36	1.00	0.86	0.93
fallacy of relevance	33	1.00	0.67	0.80
false causality	48	1.00	0.63	0.77
false dilemma	33	1.00	0.64	0.78
faulty generalization	24	1.00	0.75	0.86
intentional	27	1.00	0.63	0.77

Table 4: Per-fallacy performance from the Human-only study: number of instances (n), precision, recall, and F1-score.

sion reaching 1.00 for every type and recall remaining high overall. Perfect F1-scores (1.00) were achieved for several categories such as *appeal to emotion*, *circular reasoning*, *fallacy of extension*, *fallacy of logic*, *faulty generalization*, and *intentional*, while slightly lower values appeared for more challenging classes like *fallacy of credibility* (0.91) and *ad hominem* (0.93). Results across the 20 evaluation sets confirm this trend: most sets achieved F1-scores above 0.90, with a few dips into the 0.70–0.80 range and a single low of 0.62, indicating some variability in performance across sets. Taken together, these findings suggest that Human–AI collaboration yields near-perfect precision and generally reliable recall, producing robust F1-scores across most fallacy types and evaluation sets. Table 7 shows the performance of detection fallacy across the 20 sets, and Table 6 shows the performance across the fallacy types.

Hint Usage Analysis of hint usage revealed that participants relied most on hints for identifying non-fallacious statements (16 hints), followed by fallacies such as "false causality" and "fallacy of extension." In contrast, fallacies like "ad hominem" and "equivocation" required fewer hints, suggesting that these types were easier for participants to identify with minimal AI support. Overall, the HI approach demonstrated a significant improvement in fallacy detection performance, highlighting the potential of AI assistance in enhancing human reasoning and decision-making.

Set	n	Prec.	Rec.	F1
1	30	0.82	0.67	0.74
2	30	0.78	0.86	0.82
3	30	0.82	0.86	0.84
4	30	0.85	0.81	0.83
5	30	0.88	0.71	0.79
6	30	0.81	0.62	0.70
7	30	0.85	0.81	0.83
8	30	0.68	0.62	0.65
9	30	0.82	0.86	0.84
10	30	0.73	0.76	0.74
11	30	0.78	0.67	0.72
12	30	0.68	0.62	0.65
13	30	0.83	0.71	0.77
14	30	0.79	0.71	0.75
15	30	0.70	0.76	0.73
16	30	0.67	0.57	0.62
17	30	0.78	0.67	0.72
18	30	0.83	0.71	0.77
19	30	0.93	0.67	0.78
20	30	0.76	0.76	0.76

Table 5: Performance across 20 evaluation sets from the Human-only study: precision, recall, and F1-score.

5 Comparative Analysis of Human-Only and HI Studies

This section compares the results of the human-only fallacy detection study with the Hybrid Intelligence (HI) study, focusing on performance, additional variables such as hint usage, and the potential influence of the placebo effect. In addition, we report the results of a complementary experiment that directly evaluated large language models (LLMs) on the same fallacy detection tasks, providing a benchmark for comparison against both human-only and Human–AI studies. Together, these analyses highlight the strengths and limitations of each approach and provide insights into the role of AI in enhancing human reasoning and decision-making.

Performance The comparison between the Human-only and Human–AI reveals a clear improvement when AI support is introduced. While human alone achieve perfect precision but more variable recall across fallacy types (average F1 $\approx$ 0.84), collaboration with AI substantially boosts recall (average F1 $\approx$ 0.96), leading to more consistent performance across categories. A similar trend appears in the 20 evaluation sets: the Human-only study yields moderate stability with mean F1

Fallacy Type	n	Prec.	Rec.	F1
Ad Hominem	15	1.00	0.87	0.93
Ad Populum	10	1.00	0.90	0.95
Appeal to Emotion	7	1.00	1.00	1.00
Circular Reasoning	11	1.00	1.00	1.00
Equivocation	8	1.00	0.88	0.93
Fallacy of Credibility	12	1.00	0.83	0.91
Fallacy of Extension	10	1.00	1.00	1.00
Fallacy of Logic	12	1.00	1.00	1.00
Fallacy of Relevance	11	1.00	0.91	0.95
False Causality	16	1.00	0.94	0.97
False Dilemma	11	1.00	0.91	0.95
Faulty Generalization	8	1.00	1.00	1.00
Intentional	9	1.00	1.00	1.00

Table 6: Per-fallacy performance from the Human-AI study: number of instances (n), precision, recall, and F1-score.

Set	n	Prec.	Rec.	F1
1	10	1.00	1.00	1.00
2	10	0.88	1.00	0.93
3	10	0.78	1.00	0.88
4	10	0.88	1.00	0.93
5	10	0.88	1.00	0.93
6	10	1.00	1.00	1.00
7	10	0.88	1.00	0.93
8	10	0.83	0.71	0.77
9	10	0.70	1.00	0.82
10	10	0.70	1.00	0.82
11	10	0.67	0.57	0.62
12	10	0.88	1.00	0.93
13	10	0.88	1.00	0.93
14	10	1.00	0.86	0.92
15	10	0.88	1.00	0.93
16	10	0.86	0.86	0.86
17	10	1.00	1.00	1.00
18	10	1.00	1.00	1.00
19	10	0.71	0.71	0.71
20	10	0.88	1.00	0.93

Table 7: Performance across 20 evaluation sets from the Human-AI study: precision, recall, and F1-score.

around 0.76, whereas the Human–AI consistently reaches higher values (mean F1 $\approx$ 0.90). These results suggest that Human–AI collaboration enhances sensitivity and reliability in fallacy detection, while maintaining the already high precision observed in human judgments. Table 9 presents the average performance per set and per fallacy type for both studies (Human-only and Human–AI).

LLM-based Fallacy Detection Results. We evaluated a range of LLMs to measure their ability to distinguish between fallacious and non-fallacious arguments in our evaluation sets, using simple zero-shot prompting[4] (the full prompt is provided in Appendix). Table 8 summarizes the results in terms of precision, recall, and F1-scores. Overall, performance varies substantially across models. *GPT-4o* achieved the best overall balance, with the highest recall (0.93) and F1-score (0.92), while *GPT-3.5* yielded the highest precision (0.93) but at the cost of lower recall (0.73). *GPT-4o-mini* also performed strongly across all metrics (F1 = 0.91). Among open-source models, *DeepSeek-v3.1* reached the strongest balance (F1 = 0.84), followed by *LLaMA-3.1* and *Qwen3-235B* (F1 = 0.79–0.80). By contrast, *Claude Opus-4* showed moderate precision but notably weaker recall, resulting in the lowest F1 among the stronger contenders (0.60). The *GPT-OSS* baselines underperformed markedly, with F1-scores below 0.25. Taken together, these results suggest that frontier models such as GPT-4o, GPT-4o-mini, and GPT-3.5 provide highly reliable fallacy detection, with some open-source systems also showing competitive performance.

Compared to human performance (Table 9), the strongest LLMs reached F1-scores on par with the Human-only performance (0.76–0.84 across dimensions) but still fell short of Human–AI collaboration, which achieved up to 0.96 across fallacy types and 0.90 across evaluation sets. This gap highlights that while LLMs can approximate human judgment, they do not yet match the substantial gains observed when humans and AI work together.

Additional Variables The human-only study showed a weak positive correlation (0.325) between the length of open responses and performance, indicating that longer responses were associated with higher performance. In contrast, the HI study introduced hint usage as an additional variable, showing that non-fallacious statements required the most hints, suggesting participants found these the most difficult to assess. Overall, hints were used in 22.5% of cases, highlighting AI's role in assisting participants with more challenging fallacies.

Explanation of Findings The significant improvement in performance in the HI study may

[4]max_tokens is 128 and temperature is 0.0

LLM	Prec.	Rec.	F1
GPT-OSS 20B	0.15	0.50	0.23
GPT-OSS 120B	0.15	0.50	0.23
GPT-3.5	**0.93**	0.73	0.82
GPT-4o-mini	0.91	0.92	0.91
GPT-4o	0.92	**0.93**	**0.92**
LLaMA-3.1 70B	0.83	0.78	0.79
LLaMA-3.1 405B	0.84	0.78	0.80
Qwen3-235B	0.85	0.77	0.79
DeepSeek-v3.1	0.84	0.83	0.84
Claude Opus-4	0.84	0.60	0.60

Table 8: Fallacy detection performance of different LLMs, reported as precision, recall, and F1-scores. The best score in each column is shown in **bold**.

Eval.	Study	Prec.	Rec.	F1
Fallacy	Human-only	1.00	0.75	0.84
	Human–AI	1.00	0.93	0.96
Set	Human-only	0.80	0.73	0.76
	Human–AI	0.87	0.94	0.90

Table 9: Average precision, recall, and F1 across fallacy types and evaluation sets for Human-only vs. Human–AI.

partially be attributed to the placebo effect, where participants' belief in AI assistance positively influenced their performance (Kosch et al., 2022). The relatively low use of hints (22.5%) suggests that participants' confidence and engagement were enhanced merely by the presence of AI, even if they did not heavily rely on it for assistance. This effect likely contributed to the increased performance compared to the human-only study.

Another key finding was the difficulty in identifying non-fallacious statements, which achieved the lowest score in both studies. This aligns with (Yeh et al., 2024), who observe that expert disagreements predominantly concern whether any fallacy is present at all, with especially low agreement for the *None* (no-fallacy) class. A likely contributor is participants' personal beliefs and biases, particularly on controversial topics such as vaccination or fossil fuels, which can create judgments even when arguments are logically sound (Teneva, 2023). Together, these results highlight how demanding it is to certify the *absence* of a fallacy and the continuing need to scaffold critical thinking to decouple belief from reasoning in argument evaluation.

An additional consideration concerns the choice of LLM used in human–AI collaboration. In this study, we employed GPT-3.5 as the assisting model. While GPT-3.5 provided stable and effective support, recent evaluations (see Table 8) show that more advanced models such as GPT-4o achieve substantially higher precision and recall in fallacy detection. This suggests that the overall gains observed in our HI setting could be further amplified with stronger LLMs, highlighting the importance of carefully selecting and updating the underlying AI systems in future HI research and applications.

6 Conclusion

This study examined the effectiveness of Hybrid Intelligence (HI) in enhancing fallacy detection by comparing the performance of participants with and without AI assistance. The results demonstrated an improvement in performance when AI support was introduced, increasing from an F1-score of 0.76 in the human-only study to 0.90 in the HI study. These findings highlight the potential of HI to complement and augment human cognitive capabilities, particularly in complex domains such as argumentation analysis. However, the study also revealed persistent challenges in distinguishing logically sound arguments, which remained difficult for participants in both conditions. This suggests a need for further advancements in critical thinking and argumentation training to better differentiate between logically sound and fallacious reasoning.

The successful integration of a ChatGPT-based AI assistant into the HI framework showcased the feasibility of AI-supported fallacy detection. Nevertheless, certain limitations, such as occasional inconsistencies in AI behavior and variability in the quality of hints, indicate the necessity for further refinement in AI design and user interaction. Despite these challenges, the study provides valuable insights into the potential applications of HI systems across various fields, including education and decision-making, where enhanced cognitive support is essential. Future research should focus on expanding participant samples to include larger and more diverse populations, refining AI systems to improve reliability and usability, and exploring the long-term impact of HI on cognitive performance and reasoning tasks. By addressing these areas, the field can better harness the potential of HI to support and enhance human decision-making and problem-solving.

Acknowledgments

This work was developed while the first author was a Master's student at the University of Groningen. We thank the anonymous reviewers for their valuable feedback and constructive suggestions, which helped improve the quality of this paper.

References

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo R Simari, Matthias Thimm, and Serena Villata. 2017. Toward artificial argumentation. *AI Magazine*, 38(3):25–36.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My ai friend: How users of a social chatbot understand their human–ai friendship. *Human Communication Research*, 48(3):404–429.

Hai Dang, Lukas Mecke, Florian Lehmann, Sarah Goller, and Daniel Buschek. 2022. How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643.

André Duarte. 2024. Epidemiology of fallacies. *Argumentation*.

Andy Field and Graham Hole. 2023. *How to Design and Report Experiments*, 3 edition. SAGE Publications.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 163–172.

Pietro Goffredo, Mauricio Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112.

Pietro Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 575–581.

Kai Guo, Yuyang Zhong, Dan Li, and Samuel Kai Wah Chu. 2023. Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. *Computers & Education*, 203:104862.

Jinghui Jiang, Amanda J Karran, Constantinos K Coursaris, Pierre-Majorique Léger, and Jörg Beringer. 2022. A situation awareness perspective on human-ai interaction: Tensions and opportunities. *International Journal of Human-Computer Interaction*, 39(9):1789–1806.

Zhijing Jin, Ayush Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yi Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198.

Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer-Human Interaction*, 29(6):1–32.

Lina Markauskaite, Rebecca Marrone, Oleksandra Poquet, Simon Knight, Roberto Martinez-Maldonado, Sarah Howard, Jo Tondeur, Maarten De Laat, Simon Buckingham Shum, Dragan Gašević, and George Siemens. 2022. Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with ai? *Computers and Education: Artificial Intelligence*, 3:100056.

N Muthmainnah, PMI Seraj, and Ibrahim Oteir. 2022. Playing with ai to investigate human-computer interaction technology and improving critical thinking skills to pursue 21st century age. *Education Research International*, 2022:1–17.

Ashish Sharma, I Wei Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Zohreh Sourati, VP Venkatesh, Dhruv Deshpande, Harsh Rawlani, Filip Ilievski, Helen Sandlin, and Alain Mermoud. 2023. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418.

Elena V Teneva. 2023. Digital pseudo-identification in the post-truth era: Exploring logical fallacies in the mainstream media coverage of the covid-19 vaccines. *Social Sciences*, 12(8):457.

Frans H Van Eemeren and Bart Verheij. 2017. Argumentation theory in formal and computational perspective. In *Handbook of Formal Argumentation*, volume 1, pages 3–71.

Polyxeni Vassilakopoulou, Arve Haug, Lars M Salvesen, and Ilias O Pappas. 2022. Developing human/ai interactions for chat-based customer services: lessons learned from the norwegian government. *European Journal of Information Systems*, 32(1):10–22.

Lucie Vrbová, Katerina Jiřinova, Karel Helman, and Hana Lorencova. 2021. Do informal reasoning fallacies really shape decisions? experimental evidence. *Rationality and Society*, 33(4):448–479.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. Cocolofa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.

7 Appendix

7.1 System Prompt in Study 2

This is part[5] of the System Prompt used in 'Study 2: ChatGPT-based Hybrid Intelligence Fallacy Detection':

/system

You are an interactive assistant used for conducting a user study about AI-human interaction for fallacy detection. Your task is to collaboratively assist participants in identifying fallacies in a series of arguments.

IMPORTANT: 1. You must NOT assess or evaluate whether the participant's answers are correct or incorrect. 2. Your role is to facilitate discussion, record the participant's answers, and provide subtle guidance (if requested) without indicating correctness. 3. Decisions about the presence of fallacies should be made collaboratively, with input from both the participant and the hints you provide, if requested. 4. Every question requires a response. Short answers (e.g., "yes/no") and open explanations are mandatory and cannot be skipped. 5. Maintain a neutral tone throughout the session.

Instructions: 1. **Introduction**: Start by greeting the participant and introducing the task. 2. **Confirm Instructions**: Ask the participant to confirm that they have read and understood the instructions. 3. **Consent to Participate**: Ask the participant to provide consent to participate in the study. They must type "yes" to confirm their consent before proceeding. 4. **Pre-Task Questions**: - Ask if the participant uses ChatGPT (Yes/No). - If "Yes," follow up with: 1. "How often do you use ChatGPT?" (e.g., Daily, Weekly, Occasionally). 2. "In which areas or domains do you use ChatGPT?" - Ask: "Do you trust AI for decision-making? (Yes/No)." - Follow up with: "Why or why not?" 5. **Argument Presentation**: Present each argument one by one. 6. **Fallacy Identification**: - Ask if there is a fallacy in the argument (yes/no). Ensure they respond with "yes" or "no." If they type something else, ask them to retype their answer to make it "yes" or "no." - After the short answer, ask the participant to explain their reasoning in an open-ended way. They must provide an explanation. If the explanation is too short or unclear (e.g., "idk," "seems good"), ask for clarification or elaboration. 7. **Provide Hints**: - If the participant types "Hint," provide a subtle suggestion to help them think more critically about the argument. - After each hint, ask: - "Based on this hint, can you now provide a yes/no answer and explain your reasoning in detail?" 8. **Encouraging and Neutral**: Maintain an encouraging tone and remain neutral. Do not indicate whether their answers are correct or incorrect. 9. **Trustworthiness Assessment**: - Present the questions one by one. All questions are mandatory, and participants must provide an answer. 1. "Do you think the hints I provided were helpful overall? (Rate from 1 to 5.)" 2. "What if I told you that some of the hints I provided were incorrect? Would this change your trust in me or your answers? Why or why not?" 3. "Do you think some of the hints were incorrect on purpose? Why or why not?" 4. "Based on your experience, how likely do you think it is that my hints were accurate? (Rate from 1 to 5.)" 5. "Do you feel the hints influenced your reasoning or just confirmed what you already believed?" 10. **Session Completion**: After all arguments are completed, thank the participant for their time, ask for optional feedback, and instruct them to save and send their interaction to the study coordinator.

—

7.2 LLM Prompt for Fallacy Detection

"You are a critical thinking expert. Determine if the following argument contains a logical fallacy. If yes, reply exactly "fallacy". If no, reply exactly "no_fallacy"."

[5] The remaining part provides an example of the discussion flow.