Trust in Information in the Age of Generative AI

Using AI Personas to Evaluate Trustworthiness and Misinformation Detection

Jeremiasz Drohomirecki

Supervisors: Ujwal Gadiraju, Esra de Groot, Marije van Dalen, Shreyan Biswas

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 19, 2026

Name of the student: Jeremiasz Drohomirecki
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Esra de Groot, Marije van Dalen, Shreyan Biswas, Myrthe Tielman

Abstract

The increasing use of generative AI has had a significant impact on how people experience, interact, and interpret media. The widespread adoption of generative AI has raised concerns regarding the spread of AI-generated misinformation and its influence on the perceived trustworthiness of information. This study investigated how AI-personas representing young adults evaluated AI-generated and human-generated statements. A mixed factorial experimental design was used with three independent variables: statement truthfulness, statement source, and source label visibility. 124 AI-personas completed surveys where they were asked to evaluate short statements based on their truthfulness, confidence and trustworthiness. Mixed ANOVA was conducted to examine the effects of content source, truthfulness and labeling.

The results showed that AI-generated misinformation was not identified less accurately than human-generated misinformation. Source labeling did not significantly affect confidence in truthfulness judgments. Trustworthiness ratings were significantly influenced by both statement condition and label visibility. When source labels were hidden, AI-generated statements received higher trustworthiness ratings than human-generated statements. However, when the source labels were revealed, the trustworthiness ratings for AI-generated content were reduced, while human-made statements received higher trustworthiness scores. These findings suggest that knowledge of content origin influences the perceived trustworthiness.

1 Introduction

The recent spread of AI-generated content has had a significant impact on how young adults experience, interact, and interpret media [1]. AI generated content containing misinformation is becoming more widespread while also becoming harder to detect and more convincing [2]. It is important to understand how such content can influence people and how it affects their perception of and trust of publicly accessible information, as misinformation can shape public opinion, influence decision-making, and reduce confidence in trustworthy sources.

Recent research shows that generative AI systems can both unintentionally and deliberately produce misinformation [3]. This behavior is even more common in politically and culturally sensitive domains where objective truth is absent and facts are disputed or ambiguous. Furthermore, the detection of AI-generated misinformation has been shown to be highly context-dependent, with performance varying based on many factors such as prompting techniques, use of emotional labels, and language. Attempts at mitigating AI-generated misinformation have shown inconsistent results, even in some cases, exposing users to contradictory information when the same prompt was reused [3].

Despite recent studies aiming to improve our understanding of how AI-generated misinformation is created and mitigated, very few focused on its impact on young adults' trust in information [3]. Young adults, those being people aged 18–25 are especially at risk of exposure to AI-generated misinformation, which can affect their perception of media trustworthiness because they are among the most active users of social media platforms where AI-generated content is frequently consumed [4]. In order to learn more about this impact, we propose the following research questions:

Research questions

1. **RQ1:** How accurately can young adults distinguish between true and false content when it is AI-generated versus human-generated?
2. **RQ2:** How do perceived trustworthiness ratings differ between AI-generated and human-generated content?
3. **RQ3:** How does labeling content as AI-generated versus human-created affect perceived trustworthiness?
4. **RQ4:** How does labeling content as AI-generated versus human-generated affect distinguishing between true and false content?

This research aims to further understand the impact of AI-generated misinformation on young adults, primarily how accurately they can distinguish between true and false information and how trust differs when evaluating AI-generated information. For the purposes of this experiment, AI-personas representing a diverse range of young adults were used in place of human participants.

2 Background and Hypotheses

AI-Generated Misinformation

AI-generated misinformation can be an inaccurate, false, or deceptive statement created by generative AI systems such as ChatGPT, Claude, or Google Gemini. These tools are able to create synthetic content from human-made prompts that appears real because many of the systems are trained on large quantities of publicly available human made material. While the rapid advancement in AI technologies has offered increased efficiency, it has also created many concerns surrounding authenticity, accountability, and integrity [5].

While malicious actors may intentionally create prompts that result in generating disinformation, this is not the only way false information is spread. A common cause of untruthful information being generated is AI hallucination. This happens when LLMs respond with reasonable but false information, sometimes even fabricating quotes and sources to add credibility to misleading statements [6]. Hallucinations have also been shown to be more frequent in politically and culturally sensitive domains, suggesting that misinformation is more likely to be generated in domains where objective truth is less clear, as reported in a recent scoping review [3]. Because LLMs are trained on real-world data, they are susceptible to repeating existing false claims, as the data used during training also contains false information. These statements can be found even in trusted sources such as newspapers and

medical journals, which can make them appear more credible. Generative AI has been shown to repeat false statements found in those sources, despite those statements later being publicly corrected [7].

There have been attempts to mitigate the spread of misinformation generated by AI systems through the implementation of safeguards. These safeguards are built to prevent LLMs from generating harmful or false information, even when prompted to do so. Despite these precautions, researchers have shown that it is possible for many models to be jailbroken—rendering the safeguards ineffective. Certain models have shown greater resilience than others, but inconsistencies in their application highlight the need for further improvements to the safeguard system [8].

Trust and Trustworthiness

Trustworthiness can be understood as the perceived reliability and credibility of information. Within misinformation research, trustworthiness is a commonly discussed topic because evaluating statements or news often requires participants to rely on surrounding context. Research on information credibility has shown that evaluating the credibility of online information has become a difficult task for humans given their limited cognitive capabilities [9]. The shift from acquiring information and news from reputable and trusted entities to accessing it through social media has resulted in consumers, rather than industry experts, becoming the parties responsible for evaluating information [10].

When assessing information, individuals often rely on perceived source credibility. Previous misinformation research has shown that source credibility can be more influential on individuals' judgments of information than the actual content when the source is perceived as highly relevant [11]. While source credibility and information trustworthiness are related concepts, they describe different aspects. A statement can be perceived as trustworthy purely based on its content, regardless of the credibility of its source. For this reason, this study defines trustworthiness as the perceived reliability and credibility of a specific piece of information rather than trust in the source that created it.

Source labels introduce another factor that may influence trustworthiness judgments. Multiple studies have investigated whether labeling content as AI-generated influences perceived credibility and accuracy. These studies have not found any significant impact on credibility ratings for AI-labeled content in general cases; however, specific factors—such as prior experience and content category—did show significant effects [12, 13].

Building on previous work, this study examines trustworthiness as an attribute of the information itself while manipulating both the actual source of the statement (AI/Human) and the visibility of source labels. This enables the analysis of trustworthiness ratings under controlled conditions, creating an environment that allows for the investigation of the influencing variables—namely, the content itself, knowledge of its origin, or potentially a combination of both.

Hypotheses

Previous research has shown that AI-generated content can be difficult to distinguish from human-made content and that LLMs are capable of producing convincing misinformation [2, 3]. This suggests that users may have difficulty accurately identifying false information when it is generated by AI. Based on these findings and to address the research questions outlined in this study, the following hypotheses are proposed:

H1 (RQ1): AI-generated misinformation will be identified less accurately than human-generated misinformation.

Studies comparing AI-generated and human-generated content have found that users assign similar levels of credibility when source labels are not visible [14]. Since LLMs are capable of producing content that mimics human-generated content, trustworthiness ratings may be based primarily on the content of the statement itself.

H2 (RQ2): When the source label is not disclosed, AI-generated statements will receive trustworthiness ratings similar to those of human-generated statements.

Research evaluating the effect of AI-generated content labels has shown no significant impact on perceived credibility regardless of whether labeling is used. The presence of a source label is therefore expected to have no impact on participants' assessments of the trustworthiness of a specific statement [12].

H3 (RQ3): Statements labeled as AI-generated will receive similar levels of trustworthiness ratings to statements labeled as human-generated.

Source labels may also influence how people evaluate the truthfulness of statements. Previous studies have shown that individuals use cues such as source credibility when forming opinions about information [11]. Some research has also documented skepticism associated with AI labels, showing that headlines labeled as AI-generated reduced participants' perceived accuracy. However, this research did not find a connection between AI-labels and "False" evaluations [15]. As such the following hypothesis is proposed:

H4 (RQ4): Statements labeled as AI-generated will be evaluated with lower confidence levels than statements labeled as human-generated.

3 Methodology

Research Design

This research used a quantitative mixed-factorial experimental design to investigate how young adults evaluate misinformation and trustworthiness in AI-generated and human-made content. The experiment evaluated whether participants were able to distinguish between true and false statements, how trustworthy they perceived the information to be, and whether exposure to source labels influenced their responses.

Three independent variables were selected: statement truthfulness (true or false), statement source (AI-generated or human-generated), and source label visibility (labeled or unlabeled). The experiment used a mixed-factorial design consisting of one between-subjects factor and two within-subjects factors. The between-subjects factor was source label visibility, meaning that participants were assigned to only

one of two experimental conditions: a labeled condition or an unlabeled condition. This resulted in two separate experiments, allowing the effect of source labeling to be examined without participants experiencing both conditions. Separating the conditions reduced the risk of labeled statements influencing opinions formed about subsequent unlabeled statements. Within each experiment, statement truthfulness and statement source served as the within-subjects factors.

In this study, trustworthiness is the metric used to evaluate the perceived reliability and credibility of a specific piece of information. It does not refer to the source from which the information originated, allowing the content to be isolated from the credibility of its source. With this definition in place, RQ3 specifically examines whether disclosing source labels affects the perceived trustworthiness of statements.

Despite AI-personas being used in place of human participants for the purposes of this research, the experiment was designed to closely resemble one involving human subjects. The same datasets, variables, survey, and analysis could be used in future studies with human participants for replication and verification purposes. This would additionally provide further insight into how well AI-personas can mimic human behavior in studies related to trust evaluation.

AI-Persona Participants

Instead of using human participants, AI-personas were used for the purposes of the experiment. Each persona was created with a set of traits designed to mimic those of a young adult. These traits included AI literacy, trust in AI, and specialized knowledge rankings across multiple domains such as geography, science, and entertainment. These traits were selected because they were expected to influence participants' responses independently of the experimental variables. AI literacy may affect an individual's ability to recognize AI-generated content and evaluate its credibility, influencing both source identification and trustworthiness ratings. Trust in AI may influence the trustworthiness ratings assigned to AI-generated statements regardless of their actual truthfulness. Specialized knowledge within the statement domains (Geography and History, Science and Health, Entertainment and Literature, and Technology and Internet) may improve the accuracy of truthfulness judgments for statements related to those topics.

The use of AI-personas helped avoid several ethical concerns, such as the need for ethics committee approval and the risks associated with intentionally exposing participants to misinformation during the experiment. It is important to note that AI-personas do not represent actual human cognition or emotions. Previous research evaluating the ability of AI subjects to mimic human responses has produced mixed findings. Several studies have shown that LLMs can successfully reproduce patterns found in human survey and behavioral data [16, 17]. However, other research has identified limitations in the ability of LLMs to replicate human cognition, especially when existing cognitive psychology tasks are modified or presented in slightly different ways [18]. These findings show that, while LLMs can be used to simulate certain aspects of human behavior, caution should be exercised when generalizing their responses to real human demographics. As such, AI-personas are used only as a simulation environment for exploring potential patterns and testing the proposed methodology and experimental setup.

Used datasets

The experiment used a dataset consisting of 80 short statements. The dataset was equally divided according to two criteria: whether a statement was true or false, and the source of the information (AI or human). This resulted in four categories: true AI-generated statements, false AI-generated statements, true human-generated statements, and false human-generated statements. Based on the statements used, four knowledge domains were distinguished: Geography and History, Science and Health, Entertainment and Literature, and Technology and Internet. These domains were included to ensure that the experiment used statements from a diverse range of topics rather than focusing on a single area of expertise. Participants may have different levels of familiarity and knowledge across specific domains, which can improve their ability to correctly assess the truthfulness of statements related to those areas. By including statements from multiple domains, the study reduces the likelihood that the results are biased by topic-specific knowledge and allows the findings to be more generalizable. Additionally, in real-world settings, misinformation can be encountered across many different domains. By using this setup, the experiment creates a more realistic environment.

Human-generated statements were selected from the FEVER (Fact Extraction and VERification) dataset [19]. The dataset contains information extracted from Wikipedia, which, after modification, is classified as Supported, Refuted, and NotEnoughInfo. Supported statements were used for the true human-generated statement category, while refuted statements were used for the false human-generated statement category.

The same statements were used as examples for an LLM model (ChatGPT 5.5). The model was prompted to generate 20 pieces of truthful information and 20 pieces of misinformation in a style similar to the examples provided. This resulted in 40 AI-generated statements that mirrored the length, style, and topics of the original FEVER claims. Each statement was manually checked for correctness. During this process, one of the statements generated and labeled as true was: "The Great Wall of China is visible from Earth orbit." This statement is a well-known myth that has been disproven for many years. Following the manual review, the false statement was removed and replaced with another AI-generated statement.

Procedure

The study was divided into two experiments, with the first 250 personas assigned to Experiment A and the remaining 250 personas assigned to Experiment B.

Experiment A: Unlabeled statements

Under the unlabeled condition, participants were not informed whether a statement was AI-generated or human-generated. Each AI-persona was first initialized using a profile describing demographic and personal traits. Participants

were then presented with a single statement and asked to answer the following four questions:

1. Is the statement true or false? (Allowed answers: T/F/Unsure)
2. How confident are you in that judgment? (1–7 scale ranging from not confident at all to extremely confident)
3. Who do you think probably created this statement? (1–7 scale ranging from definitely human to definitely AI)
4. How trustworthy does this statement seem? (1–7 scale ranging from not trustworthy at all to extremely trustworthy)

Each participant evaluated five randomly selected statements. Random selection was used to reduce ordering and selection biases. Because individual statements may vary in difficulty or familiarity, randomly distributing statements across participants helps balance these effects throughout the experiment.

Experiment B: Labeled statements

Under the labeled condition, participants were informed whether a statement was AI-generated or human-generated. Each AI-persona was initialized in the same manner as in Experiment A. Participants were then presented with a single statement and asked to answer the following three questions:

1. Is the statement true or false? (Allowed answers: T/F/Unsure)
2. How confident are you in that judgment? (1–7 scale ranging from not confident at all to extremely confident)
3. How trustworthy does this statement seem? (1–7 scale ranging from not trustworthy at all to extremely trustworthy)

The perceived-source question was not included because the source label was already visible. Each participant again evaluated five randomly selected statements. At no point were misleading labels shown to participants (e.g., an AI-generated label attached to a human-generated statement).

Both experiments were repeated multiple times using different AI-personas to create a large dataset for further analysis. Each of the 80 statements appeared an equal number of times in order to maintain a balanced distribution of responses across all statement categories.

Analysis

The collected data were analyzed using mixed-design analyses of variance (ANOVA). The experimental design consisted of one between-subjects factor (source label visibility) and two within-subjects factors (statement truthfulness and statement source). Separate mixed ANOVAs were conducted for each dependent variable: truthfulness judgments, confidence ratings, and trustworthiness ratings. The effects of the independent variables were examined to determine whether statement truthfulness, statement source, and source label visibility influenced participants' responses. Statistical significance was evaluated using an alpha level of $\alpha = 0.05$. Effect sizes were reported using partial eta squared (η_p^2).

Prior to conducting the ANOVA analyses, the underlying assumptions were evaluated. The normality of the dependent variables was assessed using the Shapiro–Wilk test. Homogeneity of variances between the labeled and unlabeled groups was examined using Levene's test. As the design included repeated-measures factors, the assumption of sphericity was assessed using Mauchly's test. These diagnostic procedures were also used to evaluate the assumptions underlying the a priori power analysis conducted in G*Power [20]. All analyses were performed in Python.

Participants

An a priori power analysis was performed using G*Power 3.1 [20] to estimate the minimum sample size required for the study given the selected parameters. The design consisted of one between-subjects factor, namely source label visibility (labeled/unlabeled), and two within-subjects factors: statement truthfulness (True/False) and statement source (AI/Human).

The power analysis was performed using the *ANOVA: Repeated Measures, Within–Between Interaction* test from the F-test family. A small-to-medium effect size of $f = 0.15$ was assumed, together with a significance level of $\alpha = 0.05$ and a desired statistical power of $1 - \beta = 0.80$. The number of groups was set to two (Labeled/Unlabeled), while the number of repeated measurements was set to four, corresponding to the four statement categories. A nonsphericity correction of $\varepsilon = 1$ was assumed.

The analysis indicated a required sample size of 486 participants to achieve the desired statistical power. For this research, which used AI-personas instead of human participants, 500 personas were created and assigned to the experiments. Due to the random assignment of statements, not all participants were exposed to statements representing each of the four categories (AI-true, AI-false, human-true, and human-false). Because the mixed ANOVA required observations for all repeated-measures variables, only the 124 participants who were exposed to all four statement categories were included in the final analysis.

4 Results

Data Preparation and Assumption Testing

For the experiments, 500 AI-personas were generated. Each of them answered questions regarding 5 random statements, out of a pool of 80. This resulted in a total of 2,500 answers being collected regarding the statements. Due to the random assignment of statements, many participants were not exposed to each of the four categories of statements (AI-true, AI-false, Human true, Human false). For the mixed ANOVA analysis, only 124 participants who were exposed to all four categories were included in the analysis. Since each participant answered questions regarding 5 statements, one of the statement categories was included twice. The responses for the repeated category were aggregated for each subject.

Shapiro-Wilk tests showed that all of the dependent variables violated the assumption of normality ($p < 0.01$). However due to Shapiro-Wilk test being sensitive to small deviations on large sample sized datasets, Q-Q plots were in-

spected. The manual analysis of the plots indicated moderate deviations from normality for confidence measures and minor deviations from the distribution line for the trustworthiness scores. Accuracy did not show any normal distribution, as it showed a clear binary pattern. Due to the robustness of mixed ANOVA to violations of normality, this assumption was considered acceptable.

Levene's test indicated that the homogeneity of variance assumption was violated for accuracy ($p = 0.002$), suggesting unequal variances across conditions for this dependent variable. No evidence of violation was found for confidence ($p = 0.215$) or trustworthiness ($p = 0.104$), indicating that variances across conditions were sufficiently similar for these outcomes. Given the repeated-measures design and large sample size, the mixed ANOVA was considered sufficiently robust to proceed despite the violation observed for accuracy.

Mauchly's test showed that the assumption of sphericity was violated for all dependent variables: accuracy ($W = 0.044, p < 0.001$), confidence ($W = 0.640, p < 0.001$), and trustworthiness ($W = 0.698, p < 0.001$). As a result, Greenhouse-Geisser corrections were applied in the ANOVA analyses in order to reduce the likelihood of false positives.

Truthfulness

Table 1: Statistics for accuracy across experimental conditions

Condition	Mean	SD
Labeled - False AI	1.000	0.000
Labeled - False Human	0.966	0.126
Labeled - True AI	1.000	0.000
Labeled - True Human	0.958	0.191
Unlabeled - False AI	1.000	0.000
Unlabeled - False Human	0.892	0.262
Unlabeled - True AI	0.992	0.064
Unlabeled - True Human	0.825	0.378

Participants showed very high accuracy in identifying the truthfulness of statements across all conditions. The accuracy recorded for each condition can be seen in Table 1. AI-generated false statements had the highest accuracy, and human true statements had the lowest.

A mixed-design ANOVA with Greenhouse–Geisser correction (due to violated sphericity) revealed a significant main effect of label visibility ($F(1, 114) = 11.13$, $p = 0.001$, $\eta_p^2 = 0.089$) and a significant main effect of condition ($F(3, 342) = 9.37$, $p_{GG} < 0.001$, $\eta_p^2 = 0.076$). The interaction between label visibility and condition was also significant ($F(3, 342) = 3.09$, $p = 0.027$, $\eta_p^2 = 0.026$).

Confidence in Truthfulness

Participants' confidence ratings were high across all experimental conditions, as seen in Table 2. Mean confidence ranged from 4.79 (Unlabeled - False Human) to 5.76 (Unlabeled - True AI).

A mixed-design ANOVA with Greenhouse-Geisser correction (due to violated sphericity) revealed a non-significant main effect of label visibility ($F(1, 114) = 0.12$, $p = 0.731$, $\eta_p^2 = 0.001$) and a significant main effect of condition

Table 2: Descriptive statistics for confidence across experimental conditions

Condition	Mean	SD
Labeled - False AI	5.230	1.736
Labeled - False Human	4.976	1.193
Labeled - True AI	5.587	0.710
Labeled - True Human	4.952	0.874
Unlabeled - False AI	5.516	1.147
Unlabeled - False Human	4.787	1.318
Unlabeled - True AI	5.762	0.537
Unlabeled - True Human	5.098	0.855

($F(3, 342) = 12.80$, $p_{GG} < 0.001$, $\eta_p^2 = 0.101$). The interaction between label visibility and condition was not significant ($F(3, 342) = 1.95$, $p = 0.122$, $\eta_p^2 = 0.017$).

Trustworthiness

Table 3: Descriptive statistics for trustworthiness across experimental conditions

Condition	Mean	SD
Labeled - False AI	1.952	0.566
Labeled - False Human	3.730	1.285
Labeled - True AI	3.881	1.214
Labeled - True Human	5.008	0.738
Unlabeled - False AI	2.762	0.820
Unlabeled - False Human	3.246	0.990
Unlabeled - True AI	5.377	0.650
Unlabeled - True Human	4.623	1.178

Participants' trustworthiness ratings varied across conditions, as shown in Table 3. Overall, false AI-generated statements received the lowest trustworthiness ratings, while true AI-generated statements under the unlabeled condition received the highest ratings.

A mixed-design ANOVA with Greenhouse-Geisser correction (due to violated sphericity) revealed a significant main effect of label visibility ($F(1, 114) = 11.45$, $p < 0.001$, $\eta_p^2 = 0.091$). There was also a significant main effect of condition ($F(3, 342) = 183.42$, $p_{GG} < 0.001$, $\eta_p^2 = 0.617$). In addition, a significant interaction effect between label visibility and condition was observed ($F(3, 342) = 28.24$, $p < 0.001$, $\eta_p^2 = 0.199$).

Post-hoc Comparisons

Benferroni-corrected post-hoc comparisons revealed that the significant condition effect for accuracy was mainly caused by lower accuracy on human-generated statements. No significant difference was found between true and false AI-generated statements. For confidence ratings, only the comparison between false human-generated and true AI-generated statements remained significant after correction. Trustworthiness ratings showed significant differences between all statement categories.

5 Discussion

This study investigated how label visibility (labeled vs. unlabeled content) and statement condition (true vs. false, AI vs. human source) influence participants' judgments of accuracy, confidence, and trustworthiness. A mixed-design ANOVA was conducted to test hypotheses H1–H4 across the three dependent variables.

Overall, the results show a consistent pattern across dependent variables, with condition having a stronger and more robust effect than label visibility in most cases. However, label visibility still significantly influenced some outcomes, particularly trustworthiness and accuracy.

Truthfulness (Accuracy)

Participants showed very high levels of accuracy across all experimental conditions. Contrary to H1, AI-generated misinformation was not more difficult to identify than human-generated misinformation. The results showed accuracy scores for human-generated content were generally lower than for AI-generated content, particularly in the unlabeled condition. The mixed ANOVA revealed significant effects of both statement condition and label visibility, as well as a significant interaction between the factors.

The rejection of H1 suggests that AI-generated misinformation used in this experiment was not perceived as more convincing than human-generated misinformation. This could have been caused by the generated statements containing linguistic cues that made them easier to identify as false.

The extremely high accuracy scores for all measured conditions indicate a potential ceiling effect. Many participants correctly evaluated nearly every statement, and under multiple conditions, every single participant correctly evaluated them. This might have been caused by the statements being too short or too easily verifiable, making them trivial to assess. With the use of AI-personas for the purposes of the experiment, it is also possible that parts of the dataset used for the experiment were also part of the LLMs' training data, especially in the case of the FEVER dataset. This would cause the model not to reason about the statement's truthfulness, but instead to recall the truthfulness from the training process.

The significant effect of label visibility indicates that source information influenced truthfulness judgments. However, because accuracy remained very high regardless of condition, the practical impact of labeling appears limited.

Confidence in Truthfulness Judgments

Participants reported relatively high confidence levels across all experimental conditions, with mean confidence ratings ranging from 4.79 to 5.76 on the seven-point scale. While the ANOVA revealed a significant effect of statement condition, there was no significant effect of label visibility and no significant interaction between condition and labeling.

These findings do not support H4, which predicted that AI-generated labels would reduce participants' confidence in their evaluations. Participants appeared equally confident regardless of whether source labels were shown. Confidence appeared to be influenced more strongly by the content of the statements themselves than by information regarding their origin.

The significant condition effect indicates that confidence varied depending on whether statements were true or false and whether they were AI-generated or human-generated. Participants were generally most confident when evaluating AI-generated statements and somewhat less confident when evaluating human-generated statements. This pattern mirrors the accuracy results, where AI-generated statements were also associated with higher accuracy.

Trustworthiness and Source Labels

Trustworthiness ratings produced the strongest and most significant effects observed in the study. Significant main effects were found for both statement condition and label visibility, together with a large interaction effect between the two variables. These findings indicate that participants' perceptions of trustworthiness depended not only on the content of the statements but also on whether information about the source was available.

In opposition to H2, AI-generated and human-generated statements did not receive similar trustworthiness ratings when source labels were hidden. In the unlabeled condition, AI-generated statements were rated as significantly more trustworthy than human-generated statements. This finding differs from previous research that reported little difference in perceived credibility between AI-generated and human-generated content when source information was unavailable [14].

H3 was also not supported. The significant effect of label visibility demonstrates that revealing source information influenced trustworthiness evaluations. The descriptive statistics show that labeling reduced trustworthiness ratings for AI-generated statements while increasing ratings for human-generated statements. This pattern indicates that participants may have applied different standards when evaluating information once the source was part of the given information. Rather than relying on the content itself, participants appear to have made assumptions regarding the reliability of AI and human sources in their judgments.

The trustworthiness results show that source information matters when evaluating content. While the added labels did not significantly change accuracy, they substantially changed perceived trustworthiness. These findings highlight the importance of differentiating between belief in information truthfulness and its perceived trustworthiness.

6 Responsible Research

Reproducibility

The exact statements used in the experiment can be found in the project's repository [21], together with the prompts, code used to run the experiments, analysis scripts, and the collected responses from AI-personas. AI-personas were generated using the Llama 3.1 model with 8B parameters and Q4_K_M quantization.

AI-Personas

Despite the methodology and experimental setup being designed for a human survey, AI-personas were used instead of human participants for the purposes of the research project.

There were many advantages to running an experiment using synthetic users. As no human participants were involved, formal ethics approval for a human-subject study was not required. Many risks associated with running a study involving misinformation could therefore be avoided. There was no need to deceive human participants or expose them to false information that could have a negative impact on them. This also meant that no debriefing session was required, which otherwise would have been an essential part of the experiment. Since no personal data from human participants was collected, privacy concerns and GDPR obligations did not apply.

Although using AI-personas came with less risk compared to human subjects, it is important not to generalize the findings to real humans. Results from this study using AI-personas should be interpreted cautiously, as LLMs cannot fully represent human cognition and decision-making. Synthetic users may oversimplify the variability found in human reasoning. Since the dataset used for the experiment was publicly available, there is also a high chance of it being part of the LLMs' training data. This raises the question of whether the participants evaluated statements based on their content, or whether they were simply "recalling" data they had seen previously. AI-personas also introduce a limitation in reproducibility, which is caused by the randomness involved in their generation and responses. Even when the same LLM model is used together with shared code and datasets, the exact behavior cannot be guaranteed due to the LLM's nondeterministic nature.

External tools used

For the purposes of the study, Google Scholar was used for searching related literature. Python 3.13, together with the Pingouin, Pandas, and NumPy libraries, was used for experiment setup and analysis. ChatGPT 5.5 was used for code validation and writing assistance.

7 Conclusions and Future Work

This study investigated how AI-personas representing young adults (18–25) evaluate AI-generated and human-generated misinformation. It focused on how accurately participants could distinguish true statements from false ones, how trustworthy they found the statements to be, and the impact of source labeling on their evaluations.

The results showed that AI-generated misinformation was not more difficult to identify than human-generated misinformation. Participants showed high accuracy levels across all conditions, although this is likely due to the experiment's setup, resulting in a potential ceiling effect caused by easily verifiable statements. Statement condition had a significant effect on confidence ratings; however, there was no significant effect caused by the presence of source labels. Trustworthiness ratings showed the strongest effects in the study. The trustworthiness ratings varied significantly based on both statement condition and label visibility. When the source label was hidden, AI-generated statements were rated as significantly more trustworthy than human-generated statements. In cases where the source label was revealed to participants, AI-generated statements were rated as less trustworthy, and human-generated content was evaluated as more trustworthy. These findings suggest that knowledge of a content's origin influences perceived trustworthiness.

The limitations of this study should be considered when interpreting the results. The random assignment of statements resulted in only 124 of the 500 generated AI-personas being exposed to all four statement categories. This caused the final ANOVA analysis to include a substantially smaller sample size than originally intended. Participants also achieved extremely high levels of accuracy across all conditions, indicating a potential ceiling effect. This was most likely caused by the statements selected for the purposes of the experiment, as they were short and easy to evaluate. This reduced the ability to detect differences in truthfulness ratings between conditions.

Future research should replicate the experiment using human participants in order to study whether the patterns observed in AI-personas generalize to real young adults. This comparison could also benefit the broader research community by providing insight into the viability of using AI-personas as substitutes for humans in survey-based research. Future studies should also use a more difficult-to-evaluate dataset, for example one with longer statements such as blogs or news articles. The added complexity of such content could provide a more suitable representation of how individuals evaluate information in the real world.

References

[1] Y. Lao, N. Hirvonen, and S. Larsson. "AI and authenticity: Young people's practices of information credibility assessment of AI-generated video content". In: *Journal of Information Science* 52.3 (2025). DOI: 10.1177/01655515251330605.

[2] P. Korshunov and S. Marcel. "Subjective and Objective Evaluation of Deepfake Videos". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 2510–2514. ISBN: 978-1-7281-7605-5. DOI: 10.1109/ICASSP39728.2021.9414258.

[3] S. Park and X. Nan. "Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation". In: *AI & Society* 41.2 (2026), pp. 1501–1515. DOI: 10.1007/s00146-025-02620-3.

[4] Pew Research Center. *How Americans Use Social Media*. Accessed 2026-06-08. 2024. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/.

[5] J. Mathew and A. Narayanan. "Reimagining Truth: The Role of AI-Generated Content in Shaping Media Ethics and Audience Trust in a Post-Truth Era". In: *SJCC International Journal of Communication Research* 2.1 (2025), pp. 72–82. ISSN: 3048-9334.

[6] C. Bandara. "Hallucination as Disinformation: The Role of LLMs in Amplifying Conspiracy Theories and Fake News". In: *Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems* 14.12 (Dec. 2024), pp. 65–76. URL: https://sciencespress.com/index.php/JACAIDMS/article/view/14.

[7] B. Senekal and S. Brokensha. "Is ChatGPT a friend or foe in the war on misinformation? A South African perspective". In: *Communicare : Journal for Communication Studies in Africa* 42.2 (2023), pp. 3–16. DOI: 10.36615/jcsa.v42i2.2437. URL: https://journals.co.za/doi/abs/10.36615/jcsa.v42i2.2437.

[8] B. D. Menz et al. "Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis". In: *BMJ* 384 (2024). DOI: 10.1136/bmj-2023-078538. eprint: https://www.bmj.com/content/384/bmj-2023-078538.full.pdf. URL: https://www.bmj.com/content/384/bmj-2023-078538.

[9] G. Pasi and M. Viviani. *Information Credibility in the Social Web: Contexts, Approaches, and Open Issues*. 2020. arXiv: 2001.09473 [cs.CY]. URL: https://arxiv.org/abs/2001.09473.

[10] A. J. Flanagin and M. J. Metzger. "Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility". In: *Digital Media, Youth, and Credibility*. Ed. by M. J. Metzger and A. J. Flanagin. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: MIT Press, 2008, pp. 5–28. DOI: 10.1162/dmal.9780262562324.005.

[11] V. Mang, B. M. Fennis, and K. Epstude. "Source credibility effects in misinformation research: A review and primer". In: *advances.in/psychology* 2 (2024), e443610. DOI: 10.56296/aip00028.

[12] F. Li and Y. Yang. "Impact of Artificial Intelligence–Generated Content Labels On Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation: Web-Based, Randomized, Controlled Experiment". In: *JMIR Formative Research* 8 (2024). ISSN: 2561-326X. DOI: https://doi.org/10.2196/60024. URL: https://www.sciencedirect.com/science/article/pii/S2561326X24007443.

[13] F. Li, Y. Yang, and G. Yu. "Nudging Perceived Credibility: The Impact of AIGC Labeling on User Distinction of AI-Generated Content". In: *Emerging Media* 3.2 (2025), pp. 275–304. DOI: 10.1177/27523543251317572. eprint: https://doi.org/10.1177/27523543251317572. URL: https://doi.org/10.1177/27523543251317572.

[14] M. Huschens et al. *Do You Trust ChatGPT? – Perceived Credibility of Human and AI-Generated Content*. 2023. arXiv: 2309.02524 [cs.HC]. URL: https://arxiv.org/abs/2309.02524.

[15] S. Altay and F. Gilardi. "People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation". In: *PNAS Nexus* 3.10 (2024), pgae403. DOI: 10.1093/pnasnexus/pgae403. URL: https://doi.org/10.1093/pnasnexus/pgae403.

[16] L. P. Argyle et al. "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3 (Feb. 2023), pp. 337–351. ISSN: 1476-4989. DOI: 10.1017/pan.2023.2. URL: http://dx.doi.org/10.1017/pan.2023.2.

[17] G. Aher, R. I. Arriaga, and A. T. Kalai. *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. 2023. arXiv: 2208.10264 [cs.CL]. URL: https://arxiv.org/abs/2208.10264.

[18] M. Binz and E. Schulz. "Using cognitive psychology to understand GPT-3". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2218523120. DOI: 10.1073/pnas.2218523120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2218523120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2218523120.

[19] J. Thorne et al. "FEVER: a Large-scale Dataset for Fact Extraction and VERification". In: *NAACL-HLT*. 2018.

[20] F. Faul et al. "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences". In: *Behavior Research Methods* 39.2 (2007), pp. 175–191. DOI: 10.3758/BF03193146.

[21] J. Drohomirecki. *CSE-2026-Research-Project*. GitHub repository. 2026. URL: https://github.com/JeremiaszJD/CSE-2026-Research-Project.