Can AI and Media Literacy Guidance Improve AI-Generated Content Detection? An Intervention Study with Simulated Young Adults

Bachelor Thesis

Iustin-Nicolae Tudor[1]

Responsible Professor: Dr. Ujwal Gadiraju[1]
Supervisors: Dr. Marije van Dalen[1], Esra de Groot[1], Shreyan Biswas[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to the EEMCS Faculty of Delft University of Technology,
in Partial Fulfilment of the Requirements
for the Bachelor of Computer Science and Engineering

June 21, 2026

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

Abstract

AI-generated content has become very hard to distinguish and it has evolved into a challenge for users to judge whether the media was created by a human or by a machine. This study examines whether AI and media literacy interventions can improve the ability of AI-agent personas, prompted as young adults, to detect AI-generated texts and images. Twenty AI-agent personas completed pre- and post-intervention detection tasks across both modalities. Overall detection accuracy increased from 85.75% before the intervention to 94.25% after the intervention, with a larger improvement for image stimuli compared to text stimuli. Text detection accuracy was already high before the intervention, while image detection still showed room for improvement. The findings suggest that AI and media literacy guidance can produce measurable changes by using specific cues, but they should not be treated as direct evidence of how real young adults would respond. This study contributes an exploratory test of using AI-agent personas to evaluate intervention designs before human-participant research.

1 Introduction

As generative AI becomes more accessible, the underlying question of whether online content was created by a human or a machine has become harder to answer. In recent years, generative AI has rapidly changed the production of digital content, including text, images, audio and video [7], [22]. Recent developments in generative AI have made distinguishing AI-generated content from human-created content a very difficult task across multiple media formats, including text, images, audio and video [22]. Whilst AI-generated content can be used for harmless or comical purposes, it can also be used with malicious intent, such as using deepfakes to impersonate or damage the reputation of public figures [24], [11], spread political misinformation [8] and fake news [21] or promote scams [23], [19].

The ability to distinguish between AI-generated content and human-generated content is important. During the opinion formation process, misinformation can alter the decision-making process of an individual, influencing attitudes and decisions [27]. Thus, improving the ability to detect AI-generated content is relevant when it comes to reducing the impact of misinformation.

In this study, young adults are defined as individuals aged between 18 and 25 years old, based on the concept of emerging adulthood [3]. Young adults are an important group for this type of research, due to the fact that they are highly exposed to social media and digital content, which is the epicentre for encountering news, information and misinformation [10], [29]. This age group is also located within the process of developing the critical skills needed to make informed judgments and decisions [15]. Thus, young adults form an important reference group for research on AI and media literacy interventions, especially because such interventions have been shown to improve media knowledge, critical evaluation and resilience to misinformation [14], [13].

There are already studies that have begun examining how people, and especially young people, engage with AI-generated media [17], [18], [25]. One example would be Lao et al. [18], who studied young people's media and information literacy practices in relation to deepfakes by using qualitative interviews with participants in the range of 14-15 years old. The findings show that young people encounter deepfake content very often on social media incidentally and they usually respond to it in a casual manner, whilst also possessing a basic understanding of what a deepfake is, how they are created and what the potential risk they may pose. This study is relevant because it shows that AI-generated media is already a part of young people's everyday internet use and that research should take into consideration how young people encounter, understand and respond to this type of content in daily life.

When it comes to detecting AI-generated content, prior work has identified and compared effective methods, such as warning messages and detection tips [12], [9]. Huang and Hu [12] studied the effectiveness of warning and tip messages by conducting a two-wave online experiment regarding media literacy interventions for combating deepfakes. The findings show that technique tips have managed to improve deepfake detection, irrespective of the format, whereas warning-only messages and combined warning-and-tip messages did not have the same effect. Furthermore, Guo et al. [9] discovered that specific tips were better than general tips when it comes to detecting AI-generated visual misinformation (AIVM), incidentally making people spend more time analysing the content to confirm its authenticity. For example, one specific tip instructed participants to look for abnormal details, such as unnatural hands, inconsistent limbs, or strange background elements. This study also showed that accompanying tips with images makes it easier for people to understand the tips.

The modality of AI-generated content is also relevant for this study because detecting generated text and images can require different forms of judgment. Waltzer et al. [30] found that students and instructors found it difficult to distinguish AI-generated essays from student-written essays, showing that identifying AI-generated text can be difficult in an educational context. Similarly, Nightingale and Farid [26] showed that AI-synthesised faces can be difficult to distinguish from real faces. These findings hint towards the idea that detection difficulty can differ depending on whether the content is text-based or image-based.

Combined, these studies show that the researchers have already started to identify the problems with detecting AI-generated content, but also the potential of literacy-based interventions. However, further research is still needed on how such intervention strategies may influence detection performance in the context of young adults, particularly when applied to both AI-generated images and text. In this study, AI agents prompted as young adults are used as a controlled proxy to explore this issue, rather than direct replacements for human participants. This approach allows the intervention designs to be tested in a controlled way, for example, by varying the agent's simulated background, education, or prior media literacy. However, their responses should not be treated

as direct evidence of how real young adults would respond.

This paper aims to answer the following research questions:

RQ1: To what extent do AI and media literacy interventions improve AI agents' ability to distinguish AI-generated content from human-created content when prompted as young adults?

RQ2: How does the modality of AI-generated content influence AI agents' ability to distinguish AI-generated content from human-created content?

To address these questions, the study used a controlled experiment to examine whether such interventions improve the agent's accuracy in distinguishing AI-generated content from human-created content. The study focuses on two different modalities, text and images, in order to compare which type of AI-generated content is more difficult for the agents to identify. This comparison is relevant here because detection cues and literacy strategies can be different across modalities, meaning that the effectiveness of the intervention for images may not be equal to the intervention for text.

2 Background and Related Work

2.1 Key Concepts and Study Context

In this study, it is important to mention that the AI agents are not going to be treated as outright replacements for human participants, but as simulated respondents that will be used to explore whether a controlled intervention design will be able to produce measurable changes when it comes to detection behaviour. Furthermore, young adults are understood as individuals aged 18-25. This age range is used to represent active users in the digital world who are likely to encounter AI-generated content in everyday online environments.

AI-generated content refers to digital material, such as text or images, that is either produced or significantly altered by a generative AI system. In the context of this study, detection refers to the task of judging whether a piece of content is AI-generated or human-created. Detection accuracy is measured by assessing the correctness of the judgment.

The intervention used in this study is understood as an AI and media literacy intervention because it helps in providing guidance that is intended to support more critical thinking and active evaluation of AI-generated content [20]. Moreover, this study also considers confidence, which refers to how certain the agent reports being when it comes to its judgment and it also takes into account perceived difficulty, which refers to how hard an agent reports that the classification task is [16]. These concepts define the main outcomes used to evaluate whether the intervention changes the agents' detection behaviour.

2.2 AI-Generated Content Detection and Media Literacy Interventions

Previous research has shown that very often, people struggle to distinguish AI-generated or manipulated content from authentic content [5], [16], [26], [30]. One example here would be Diel et al. [5], which shows that the human deepfake detection is a very difficult task, and sometimes performance is close to chance. This directly suggests that detection is not a straightforward task for ordinary users. Also, people may report confidence in their judgments even when their accuracy is limited [16]. Because of this, the study not only measures whether agents are correct, but also the difficulty they perceive the task to be and the confidence in their judgments. This work, therefore, motivates the study's focus on detection accuracy, confidence and perceived difficulty as outcome measures.

Another line of research focuses on whether media literacy interventions are successful when it comes to improving the detection of such content. Prior work suggests that the format of the intervention has a very big significance. [12] discovered that the technique-based tips were more effective compared to warning-only messages. Similarly, [9] found that specific tips accompanied by visual examples were more useful than general tips accompanied by visual examples when participants evaluated AI-generated misinformation. The present study, therefore, uses a guidance-based intervention, where AI agents, prompted as young adults, will receive specific cues for evaluating AI-generated content rather than only being warned that the content may be AI-generated. This prior work forms the basis for the intervention design used in the present study.

Research on young people's encounters with AI-generated media also helps to understand the broader motivation of this study [18], [17], [25]. Lao et al. show that young people encounter deepfake content on social media very often and they react to it in a casual manner, since it is seen as something relatively normal [18]. Even if this study is not tested on young adults directly, it supports the idea that AI-generated media has become a part of ordinary digital environments that people can find on the internet without even searching for it and supports the idea that literacy-based responses remain relevant. Looking at the current study, this human-centred concern is indirectly examined by using AI agents prompted as young adults. This literature motivates the young-adult framing of the study, while the use of AI-agent personas remains an exploratory simulation rather than a direct study of young adults.

2.3 AI Agents as Simulated Respondents and Research Gap

Recent work has also explored the usage of large language models (LLMs) as simulated participants [1], [2]. Aher et al. show that LLMs can be used to simulate multiple participants in controlled experimental settings [1], while Argyle et al. argue that language models can approximate responses from specific population groups when conditioned with demographic information, but only sometimes does it do so accurately [2]. This approach, however, has important limitations that need to be taken into perspective [4], [31]. Bisbee et al. warn that the synthetic responses can diverge from real human survey data and they change across model versions [4], while Wang et al. show that LLMs can misrepresent or flatten identity groups [31]. Thus, the current study treats AI agents prompted to represent young adults as

exploratory simulated respondents rather than direct substitutes for human participants. This literature complements the present study by supporting the use of AI-agent personas as an exploratory method, while also motivating caution in interpreting their responses.

Together, the previous studies show that the AI-generated content can often be difficult to detect [5], [16], [26], [30], that specific literacy interventions may help in improving detection [12], [9], [13] and that LLM-based agents can be useful for controlled experiment research [1], [2]. However, further research is still needed on how media literacy interventions may influence AI-generated content detection for young adults, particularly across different content modalities. This study thus addresses the gap by using AI agents prompted as young adults as exploratory simulated respondents, examining whether such an intervention can change the detection accuracy, confidence and perceived difficulty across different content modalities.

2.4 Hypotheses

The hypotheses are structured in accordance with the two research questions of the study and are linked to previous research on AI-generated content detection, media literacy interventions and the relationship between confidence and detection accuracy.

H1: Intervention effect. Previous research has shown that AI-generated or manipulated content can be difficult to detect [5], [16], [26], [30], while specific media literacy tips can improve detection performance and are most effective [12], [9]. Therefore, this study expects the intervention to improve detection accuracy.

- H1: AI agents prompted as young adults will show higher detection accuracy after the intervention than before the intervention.

H2: Modality differences. Since previous research has examined the detection of AI-generated content across different modalities, such as text, images, audio and video [8], [26], [30], this study is analysing whether the detection outcomes are different based on image and text stimuli.

- H2: Detection accuracy, perceived difficulty and confidence in judgement will differ between text and image stimuli.

H3: Confidence and accuracy. Previous research suggests that people can be confident in their judgments, even if their detection accuracy is limited [16]. Thus, this study examines whether confidence becomes more closely aligned with accuracy after the intervention.

- H3: AI agents' reported confidence will be more closely aligned with detection accuracy after the intervention than before the intervention.

3 Methodology

This study uses a controlled AI-agent intervention design to examine whether AI and media literacy guidance can improve the rate of detection of AI-generated content. The experiment follows a within-subject, repeated-measures structure, such that each AI-agent persona completes the same general task before and after receiving an intervention. Thus, the persona will act as its own comparison point. The within-subject design was chosen over a between-subject design because it allows the comparison of performance changes within the same persona, reducing the influence of differences between simulated personas, making the effect of the intervention easier to isolate. The study focuses on three main outcomes: detection accuracy, confidence and perceived difficulty.

3.1 Experimental Set-up

Variables

The main independent variable in this study is the intervention stage, which has two levels, pre-intervention and post-intervention. This variable is used to examine whether exposure to AI and media literacy guidance changes detection performance. The second independent variable is modality, which also has two levels, text and image. This makes it possible to compare the effects of the intervention depending on the type of content that is being evaluated.

The dependent variables are detection judgement, confidence and perceived difficulty. Detection judgement is used to calculate detection accuracy, which is the main outcome measure in AI-generated content detection research [16], [26], [30]. Confidence is included because previous studies found that confidence and detection accuracy do not always align, where participants may express high confidence, even though the detection performance is limited [16]. Perceived difficulty is included to find additional insight for understanding how challenging the detection task is perceived to be. The exact response format was created for this study, but it is based on these commonly used outcome measures. For each stimulus, the AI-agent persona judges whether the content is AI-generated or human-created, after which the judgment is compared with the known answers in order to calculate detection accuracy. Confidence was measured on a 1-7 Likert scale, where 1 means "not confident at all" and 7 means "very high confidence". Perceived difficulty is also measured on a 1-7 Likert scale, where 1 means "very easy" and 7 means "very difficult". The short reasoning response is collected as a qualitative explanation of the judgment.

For the confounding variables, several persona attributes are taken into account. These are AI literacy, prior AI use, online content scepticism, attention to detail and prior exposure to misinformation. Even though these characteristics are not the main independent variables, they may influence how each persona evaluates the stimuli. This is because prior work suggests that detection can be affected by users' knowledge of AI-generated content, confidence in their own judgments and the extent to which they analyse specific cues or suspicious details [20], [30], [9]. Methodological factors may also influence the results. Since all of the personas are implemented using the same underlying model (GPT-5.5 Extended Thinking), model-specific tendencies, biases, or detection capabilities may affect performance regardless of the intervention. The study used the same model, prompting structure and experimental procedure for all personas throughout this study to reduce the variation from this source.

Participants

The participants in this study are AI-agent personas rather than human respondents. A total of 20 personas are used. Each persona represents a simulated young adult between the ages of 18 and 25. The personas were generated using a persona-generation endpoint and included demographic information alongside five persona attributes (AI literacy, prior AI use, online content scepticism, attention to detail and prior exposure to misinformation). The values for these attributes were assigned randomly from the predefined options (low, medium, high), each with its own extended description (see Appendix A.1). The full set of generated persona profiles is available in a public GitHub repository.[1]

A priori power analysis was conducted in G*Power for the main intervention effect and the difference between text and image improvement. In both analyses, a matched-pairs t-test, two-tailed testing, medium effect size $d_z = 0.5$, an alpha level of .05 and statistical power of .80 were used. The recommended sample size was 34 personas. However, the final study uses 20 personas due to practical limitations of running the experiment manually in ChatGPT, especially the repeated image-upload requirements and file-upload limits. Running more personas would be unfeasible, taking into account the available time for conducting the experiment. Because the achieved sample size was below the recommended sample size, statistically significant findings are treated as exploratory rather than definitive. Each persona is run in a separate ChatGPT conversation using GPT-5.5 Extended Thinking mode, whilst also preserving the same persona profile throughout the full experimental session. The study should thus be interpreted as exploratory and limited in statistical power.

Tools and Materials

The experiment is conducted using ChatGPT in GPT-5.5 Extended Thinking mode. Each persona receives the same standardised prompt, where only the persona JSON is different between sessions. The prompt explains the task structure, response format and confidence and difficulty scales.

The text stimuli are selected from the RAID dataset [6]. In total, 20 text samples are used, which consist of 10 AI-generated and 10 human-created texts (see Appendix C.2). The AI-generated texts include outputs from models such as GPT-4, LLaMA-Chat, Mistral-Chat and Cohere-Chat. The text stimuli are divided into pre-intervention and post-intervention blocks, where each block contains 5 AI-generated and 5 human-created texts and in both blocks the human and AI categories include 2 Reddit posts, 1 wiki text, 1 abstract and 1 news text.

The image stimuli are selected from the Defactify/MS COCOAI dataset [28]. In total, there are 20 images used, 10 AI-generated and 10 human-created images (see Appendix C.1). The AI-generated images include outputs from models such as Midjourney v6, DALL-E 3 and Stable Diffusion 3. The human-created images are the real photographs included in the dataset. All the selected images come from the test folder and they include animals, people doing everyday tasks, objects and landscapes.

[1] https://github.com/IronBeagle1/Personas

The text intervention consists of specific tips and written guidance about possible indicators of AI-generated text, such as vocabulary patterns, grammar, sentence structure, formatting, tone, introductions, conclusions, specificity, factuality and originality (see Appendix B.2). The image intervention, symmetrically, consists of specific tips and written guidance about indicators such as anatomical implausibilities, stylistic artefacts, functional implausibilities, violations of physics and social or contextual implausibilities (see Appendix B.1).

Evaluation Method

Detection accuracy is calculated by comparing each persona's judgment with the known truth, where correct responses are coded as 1 and incorrect responses as 0. Accuracy is calculated separately for pre-intervention text/image and post-intervention text/image. Confidence and perceived difficulty are analysed using a Likert scale with 1-7 ratings provided after each stimulus. A short reasoning response was also collected, but it will only be kept for qualitative interpretation, whereas the main quantitative analysis is based on accuracy, confidence and difficulty. This reasoning is measured through the initial prompt field: "reasoning: one short sentence explaining the judgement."

3.2 Procedure

Each persona completed the experiment in its own separate ChatGPT chat. Using the same chat for the full session allowed the persona profile and intervention exposure to remain continuous across the task, similar to how the same participant would complete multiple stages in a repeated-measures intervention design. First, each persona completed 10 pre-intervention text judgements, followed by the text intervention, and then completed 10 post-intervention text judgements. After this, each persona completed 10 pre-intervention image judgments, then the image intervention was provided, followed by the 10 post-intervention image judgments. The selected stimuli were randomly ordered within each block and also arranged such that in each block, there were 5 AI-generated and 5 human-created stimuli. This modality order was chosen for practical reasons, due to the fact that running text blocks first made it possible to create and complete persona-based conversations more efficiently before the image-upload stage, which was more time-consuming and limited due to the ChatGPT plan used. However, possible modality-order effects are not to be fully ruled out and this is treated as a limitation of the study.

At the end of each session, a final prompt was used to collect the responses in one structured JSON object (see Appendix A.2). If at any point ChatGPT produced invalid output that did not follow the required JSON format or answer format, the chat was discarded and a new persona was generated.

3.3 Data Analysis

The collected data is organised such that each row represents one response to one stimulus. Each row includes persona ID, persona attributes, stimulus ID, modality, intervention stage, judgement, the known truth label, correctness, confidence, difficulty and reasoning. Descriptive statistics are calculated for each modality and intervention stage.

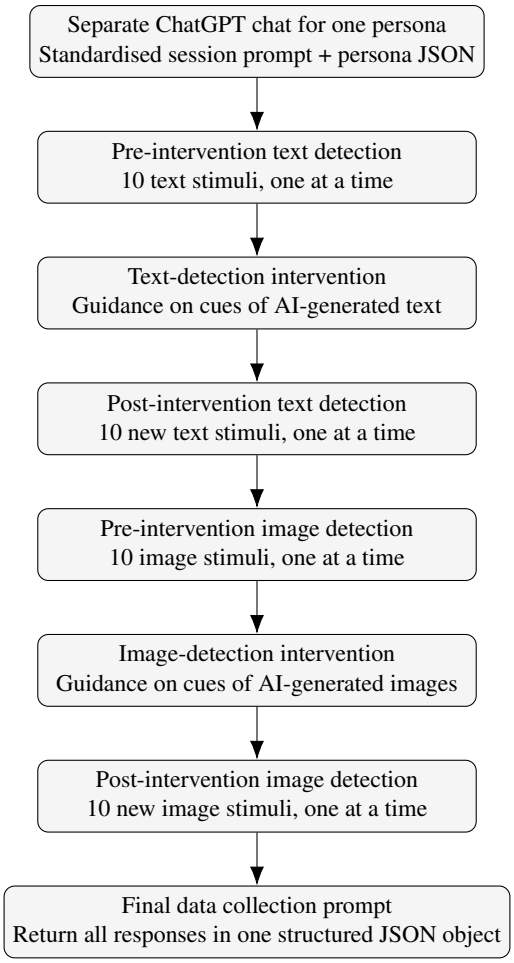


Figure 1: Overview of the experimental procedure followed by each AI-agent persona.

Note. Within each block, stimuli were shown in random order and included 5 AI-generated and 5 human-created items. If the response format was invalid or the required JSON output was not followed, the chat was discarded and a new persona was generated.

The main analysis focuses on comparing pre- and post-intervention accuracy within each persona. Since the same personas are used for completing all stages, paired-samples t-tests were used for text accuracy and image accuracy. A further paired comparison is used to examine whether improvement is different based on modality. Confidence and perceived difficulty are compared descriptively before and after the interventions, and modality differences in these measures are tested using persona-level means.

For the paired-samples t-tests, each persona contributed one mean score per relevant condition and outcome. The tests were therefore conducted on persona-level means rather than on the 800 individual judgements as independent observations. To examine H3, mean confidence was also descriptively compared between correct and incorrect judgments before and after the intervention.

4 Results

This section presents the quantitative findings of the AI-agent detection experiment. The analysis is based on a total of 800 judgements in total, where 400 are text judgements and 400 are image judgements. For each modality, 200 judgements have been collected before the intervention and 200 after the intervention. The results are reported descriptively and statistically, while a more descriptive interpretation in relation to the research questions and the hypotheses is addressed in the Discussion section.

4.1 Detection accuracy

Table 1 shows the results by modality and intervention stage. Across both modalities, the overall detection accuracy increased after the intervention. Before the intervention, 343 out of 400 judgements were correct, corresponding to an overall accuracy of 85.75%. After the intervention, 377 out of 400 judgements were correct, corresponding to an overall accuracy of 94.25%. A paired-samples t-test showed that this overall increase was statistically significant, $t(19) = 8.23$, $p < .001$, $d_z = 1.84$.

When it comes to the text stimuli, detection accuracy was very high before the intervention. This is observable since 189 out of 200 judgements were correct, achieving an accuracy of 94.5% in the pre-intervention text condition. In the post-intervention text condition, there was an increase, where 198 out 200 judgements were correct, resulting in an accuracy of 99.0%. This increase was statistically significant $t(19) = 3.33$, $p = 0.004$, $d_z = 0.74$.

For the image stimuli, detection accuracy was not as high, both in pre- and post-intervention, but it increased from 77.0% to 89.5%. In the pre-intervention image condition, 154 out of 200 judgments were correct, whilst in the post-intervention condition, 179 out of 200 judgments were correct. This increase was also statistically significant $t(19) = 8.75$, $p < 0.001$, $d_z = 1.96$.

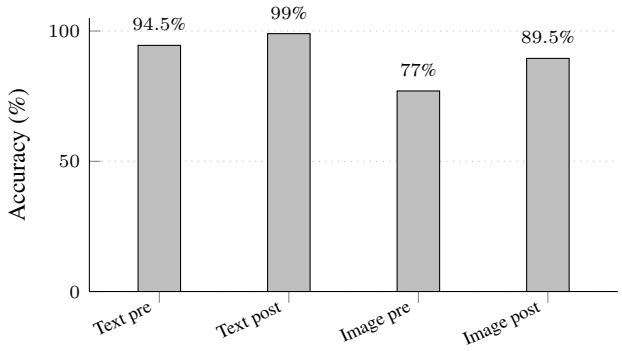


Figure 2: Detection accuracy for text and image stimuli before and after the intervention. Accuracy increased after the intervention for both modalities, with a larger improvement for image stimuli.

4.2 Differences Between Text and Image Judgements

Accuracy differed between the text and image stimuli. Across both pre- and post-intervention, conditions combined, the text

Table 1: Detection accuracy, confidence, difficulty, and error types by modality and intervention stage. Confidence and difficulty were measured on a 1–7 scale.

Condition	N	**Accuracy**	**Confidence**	**Difficulty**	**Error type**	
					AI as human	**Human as AI**
Text pre-intervention	200	94.5%	5.27	3.64	10	1
Text post-intervention	200	99.0%	5.66	3.31	0	2
Image pre-intervention	200	77.0%	5.15	3.75	44	2
Image post-intervention	200	89.5%	5.45	3.52	21	0

stimuli had an overall accuracy of 96.75%, while image stimuli had an overall accuracy of 83.25%. A paired comparison showed that image accuracy was lower than text accuracy $t(19) = -13.08$, $p < 0.001$, $d_z = -2.92$.

Confidence and perceived difficulty were also compared between text and image stimuli using persona-level means across both intervention stages. Mean confidence was higher for text stimuli ($M = 5.46$) than for image stimuli ($M = 5.30$). This difference was statistically significant, $t(19) = 4.16$, $p < 0.001$, $d_z = 0.93$. Mean perceived difficulty was higher for image stimuli ($M = 3.63$) than for text stimuli ($M = 3.47$). This difference was also statistically significant, $t(19) = 2.80$, $p = 0.011$, $d_z = 0.63$.

The improvement also differed between modalities. Text accuracy increased by 4.5 percentage points, from 94.5% to 99.0%, whilst image accuracy increased by 12.5 percentage points, from 77.0% to 89.5%. A paired comparison showed that the improvement was larger for image stimuli than text stimuli $t(19) = 4.29$, $p < 0.001$, $d_z = 0.96$.

4.3 Error patterns

Out of the 800 judgements, only 80 were classification errors. Most errors involved AI-generated content being classified as human-created. They represented 75 out of 80, whilst the remaining 5 were human-created classified as AI-generated.

The highest number of errors occurred in the pre-intervention image condition. In this condition, 44 AI-generated images were classified as human-created, whereas only 2 human-created images were classified as AI-generated. After the image intervention, the number of AI-generated images mistaken as human-created has decreased to 21. For the text stimuli, AI-generated texts were mistaken as human-created 10 times pre-intervention and 0 times post-intervention. The number of human-created texts mistaken as AI-generated was low in both conditions, with 1 before the intervention and 2 after the intervention.

4.4 Confidence and Perceived Difficulty

Mean confidence ratings increased after the intervention. Across both modalities, mean confidence increased from 5.21 to 5.55. For the text stimuli, mean confidence went from 5.27 to 5.66 after the intervention. For the image stimuli, the mean confidence increased from 5.15 to 5.45 after the intervention.

Mean perceived difficulty decreased after the intervention. Across both modalities, mean difficulty decreased from 3.69 to 3.42 after the intervention. For the text stimuli, the mean difficulty went from 3.64 to 3.31 after the intervention. For the image stimuli, the mean difficulty decreased from 3.75 to 3.52 after the intervention.

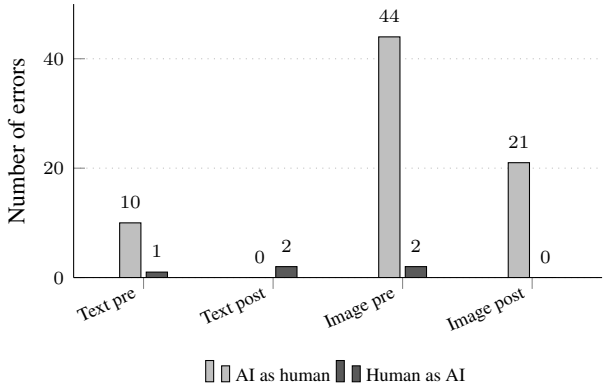


Figure 3: Misclassification types for text and image stimuli before and after the intervention. Most errors involved AI-generated content being classified as human-created, especially for image stimuli before the intervention.

Confidence was also compared between correct and incorrect judgments before and after the intervention. Before the intervention, correct judgements had a mean confidence of 5.26, while incorrect judgements had a mean confidence of 4.89, a difference of 0.36. After the intervention, correct judgements had a mean confidence of 5.58, while incorrect judgements had a mean confidence of 5.17, a difference of 0.40. This shows that correct judgements were slightly more confident than incorrect judgements in both stages, but the difference changed only slightly after the intervention.

4.5 Reasoning Responses

A total of 800 reasoning responses have also been recorded for qualitative interpretation. These responses were not used in the quantitative analysis. Instead, they were retained only to support the qualitative discussion of how personas justified their judgements and to check if the intervention cues were used.

5 Discussion

The results show three main patterns. First, detection accuracy was higher after the intervention across both modalities. Second, text stimuli were classified more accurately

than image stimuli. Third, most classification errors were AI-generated content being mistaken for human-created content, especially in the image condition before the intervention.

5.1 Interpretation of Results

This study examined whether AI and media literacy interventions can help AI agents, prompted as young adults, distinguish AI-generated content from human-created content. The results show that the detection accuracy improved after the intervention for both modalities. Because the same personas completed the task before and after the intervention, this comparison suggests that the guidance changed the agents' detection behaviour in this controlled setting.

The findings support H1, where a higher detection accuracy was expected after the intervention. Overall accuracy went from 85.75% to 94.25% after the intervention. For text stimuli, the accuracy increased from 94.5% to 99.0%, whilst for image stimuli it increased from 77.0% to 89.5%. The findings align with previous research, showing that media literacy interventions, using specific tips, can improve detection accuracy [12], [9]. In this study, the intervention did not warn that the content can be AI-generated. Instead, it gave concrete cues, which may explain why the post-intervention judgements were more accurate. Therefore, H1 is accepted, as detection accuracy increased after the intervention across both modalities.

The findings also support H2, which expected differences between the image and text stimuli. Text stimuli were more accurately classified compared to the image stimuli overall, showing that image detection was more difficult for the AI agents. One explanation for this is that text accuracy was already much higher before the intervention, leaving very little room for improvement. The model may have had a stronger baseline ability for text-based detection than image-based detection. In contrast, image accuracy was lower before the intervention, allowing a larger improvement after it. This fits with previous research that shows AI-generated images can be difficult to distinguish from real images [26]. Confidence was also higher for text stimuli than for image stimuli, while perceived difficulty was greater for image stimuli than for text stimuli. Therefore, H2 is accepted, as detection accuracy, confidence, and perceived difficulty differed between text and image stimuli.

From the error patterns, it could be observed that most mistakes involved AI-generated content being misclassified as human-created content. This was very clear for images, where 44 AI-generated images had been mistaken before the intervention, which decreased to 21 after the intervention. For the text stimuli, AI-generated texts were mistaken only 10 times before the intervention and 0 times after. Thus, it seems like the main challenge was not a balanced confusion when it came to both categories, but rather failure to identify AI-generated stimuli as artificial.

When it comes to H3, there was only limited support. Confidence increased after the intervention, while perceived difficulty decreased. However, confidence did not clearly become better calibrated with correctness. Correct judgments had a mean confidence of 5.26 before the intervention and 5.58 after the intervention. Incorrect judgments had a mean confidence of 4.89 before the intervention and 5.17 after the intervention. This means that confidence was slightly higher for correct than incorrect judgements in both stages, but incorrect judgements still received relatively high confidence ratings. The difference between confidence for correct and incorrect judgements changed only slightly, from 0.36 before the intervention to 0.40 after the intervention. Thus, even though confidence and accuracy increased after the intervention, the results do not show a clear improvement in confidence calibration. Therefore, H3 was not clearly supported.

The short reasoning responses were used only as supplementary qualitative material to contextualise the quantitative findings. They have not been analysed using a full thematic coding procedure. Thus, the observations drawn should only be treated as exploratory. Overall, the responses suggested that after the intervention, agents often justified their judgements by referring to detection cues that were part of the interventions, such as visual inconsistencies, anatomical details, object plausibility or textual style. Before the intervention, judgements appeared to rely on broader impressions of what "felt" realistic, natural, AI-generated or human-created. The observations are consistent with the quantitative findings and suggest that the interventions had an immediate effect on how agents justified their judgements.

Overall, the findings suggest that AI and media literacy guidance can improve detection performance in the case of AI-agent personas prompted as young adults. However, the results should not be seen as direct evidence of how real young adults would respond and adapt, but interpreted as evidence from a controlled AI-agent simulation.

5.2 Implications

The first implication of this study is that AI and media literacy guidance can influence how the AI-agent personas evaluate the stimuli. The intervention appeared to improve performance after the agents received specific cues regarding textual patterns, visual inconsistencies and contextual implausibilities. This suggests that cue-based interventions can be useful when testing whether detection behaviour can change after the guidance was introduced.

The second implication is that modality matters. The improvement was larger for the image stimuli, which suggests that, in the case of AI agents, image-based detection can benefit more from guidance using specific cues. However, accuracy for the text stimuli was already very high before the intervention, which created a ceiling effect. The accuracy was also higher overall for text stimuli in comparison to image stimuli, suggesting that AI-agents using the ChatGPT 5.5 Extended Thinking model have a better base proficiency when it comes to detecting and analysing text. This was also reflected in the confidence and perceived difficulty ratings, since text stimuli received higher confidence ratings, while image stimuli were perceived as more difficult. Thus, modality influenced not only detection accuracy, but also how certain the agents were and how difficult they perceived the task to be.

The third implication is that confidence should be interpreted carefully. Although confidence increased after the intervention, it did not clearly become better aligned with correctness. Thus, confidence ratings should not be treated as

reliable indicators of accurate detection on their own.

The final implication concerns the use of AI agents as exploratory research tools. These results suggest that AI-agents can be used to explore whether an intervention design can produce measurable changes. However, this also comes with a caveat, given that accuracy was already very high before the intervention, compared to previous research showing accuracy is close to chance, especially before an intervention [16], [5], [26]. Thus, AI agents should not be treated as accurately representing the way real young adults would respond and adapt based on the intervention.

5.3 Limitations

The main limitation in this study was that it used AI-agent personas rather than real humans. This was necessary and not a design choice due to not obtaining the HREC approval for human-participant research. Even though the design made it possible to test the intervention in an exploratory way, the findings cannot be generalised and linked to real young adults.

A second limitation is that all personas were implemented using the same underlying model, GPT-5.5 Extended Thinking. This was due to practical and resource constraints, as no alternative AI-agent system with comparable image upload limit rates was available for conducting repeated image-based detection tasks. Thus, the findings may partly also reflect model-specific behaviour, knowledge biases or detection capabilities, even with the variation of the given personas.

The third limitation is the sample size, caused by the file upload limit constraints of the available resources. Thus, the final study used 20 personas, even though a priori analysis recommended 34 personas. This reduced the statistical power of the study, which means that the findings should be interpreted as exploratory.

A fourth limitation is the fixed modality order. All personas went through the text-based setup before the image-based setup. Thus, possible order effects cannot be ruled out. A further limitation is that each persona completed all stages in one continuous chat. While this preserved persona continuity, later responses may have been influenced by accumulated context from earlier tasks, meaning that carry-over effects cannot be fully ruled out. The last limitation is the restricted model access. The study relied on ChatGPT as the available multimodal system. Therefore, it was not possible to compare whether other LLMs would respond differently to the same intervention.

5.4 Future Work

This study used AI-agent personas prompted as young adults to test whether an AI and media literacy intervention can improve the detection of AI-generated text and images. Because real human participants were not included, the next most important step is to test the same intervention format with young adults. This would show whether the observed improvement of the AI-agent personas can also appear in human judgment behaviour.

Future research should also compare AI-agent responses with human responses directly. Both groups could complete pre-intervention and post-intervention tasks using the same stimuli. Thus, it could further clarify whether AI agents can be useful as exploratory proxies. Also, the usage of multiple AI models could be very useful in order to determine whether the intervention effect is specific to one model. This could also show whether similar patterns appear across different LLMs.

In the future, confidence calibration should be examined in more detail. In this study, the confidence increased after the intervention. However, the incorrect judgments still received a relatively high confidence rating. Future studies could use calibration metrics to see if participants become better at knowing when they might be wrong.

6 Conclusions

This study examined how AI and media literacy interventions can help AI agents, prompted as young adults, detect AI-generated content. It investigated two research questions, determining whether such interventions improve the agents' ability to distinguish AI-generated content from human-created content and whether detection performance differs between two modalities, text and image.

The results show that the intervention improved detection performance in a controlled AI-agent setting. In relation to RQ1, accuracy increased after the intervention for both image and text detection tasks, suggesting that cue-based AI and media literacy guidance can help AI-agent personas make detection judgments more accurate. In relation to RQ2, modality influenced detection performance, confidence ratings and perceived difficulty ratings. Text stimuli were detected more accurately overall, had a higher overall confidence and a lower overall difficulty rating. Overall, H1 and H2 were accepted, while H3 was not clearly supported because confidence did not show a clear improvement in alignment with correctness after the intervention.

The study contributes by showing that AI-agent personas can be used as exploratory tools for testing intervention designs. This can be done before conducting human-participant research. However, the findings should not be treated as direct evidence of how real young adults would respond. Instead, they do show that AI and media literacy interventions using specific cues can produce measurable changes in AI-agent judgement behaviour. This also highlights the need for future validation with human participants.

7 Responsible Research

This study did not collect data from real human participants due to not obtaining HREC approval. Thus, it was conducted using AI-agent personas prompted as young adults. This avoided the collection of personal data from real, human participants, whilst also limiting the extent to which the findings can be generalised to young adults.

The experiment was made with a focus on reproducibility. It used the same task structure, response format, intervention text and stimulus blocks across all personas. The full prompts, intervention materials, text samples and image stimuli used in the experiment are provided in the appendices. The selected image stimuli are included only as part of the thesis documentation of the experiment, in line with

right-to-quote guidance for academic use, and were not redistributed separately, while the original dataset is cited as the source material. The full set of generated persona profiles is also made available in a public GitHub repository to support transparency and reproducibility.[2] The text and image stimuli have been selected from existing and readily accessible datasets and the responses have been stored in a structured format for analysis. However, the full reproducibility is limited due to the variance of LLM outputs across conversations and the unavailability of generating the persona demographics in the same way.

The researcher remained responsible for the study design, stimuli selection, data handling, analysis and interpretation of the results. Because the personas include demographic attributes, the results are not interpreted as evidence about real demographic groups or young adults in general. The limitations of using AI-agent personas instead of real humans have been reported throughout the paper in order to avoid presenting the findings as direct evidence of human behaviour. Some sentences were refined for clarity using ChatGPT and Grammarly.

A Prompt Design

A.1 Initial Persona and Session Prompt

Initial Persona and Session Prompt

```
You are participating in an AI-generated
    content detection study as one simulated
    young-adult participant.

Use the persona below as your participant
    profile for the entire session. Respond
    consistently with this persona, including
    their demographic information and listed
    attributes. Do not answer as ChatGPT, a
    researcher, or an external evaluator. It is
    highly important to try and simulate this
    persona to the best of your abilities and
    take account the persona attributes every
    time you answer.

PERSONA:
"""
{
  "persona_id": "persona_3e9476248997",
  "age": 19,
  "gender": "woman",
  "sex": "Male",
  "race": "White",
  "ethnicity": "Mexican American",
  "country": "CA",
  "state": "PA",
  "education": "vocational_training",
  "student_status": "undergraduate",
  "ai_literacy_level": "Low: Minimal
    understanding of generative AI systems.
    Rarely recognises common AI-generated
    patterns and has limited awareness of how
    AI content is produced.",
  "prior_ai_use": "Occasional: Uses AI tools
    periodically for tasks such as writing
    assistance, brainstorming, image
    generation, or information gathering.",
  "online_content_skepticism": "Low: Generally
    accepts online content at face value and
    rarely questions authenticity unless
    obvious issues are present.",
  "attention_to_detail": "High: Carefully
    analyses wording, structure, visual
    elements, and contextual consistency before
    making a judgement.",
  "prior_exposure_to_misinformation": "High:
    Extensive prior exposure to misinformation,
    manipulated media, online scams, or
    educational discussions about deceptive
    digital content."
}
"""

Session structure:
This session will be completed in the following
    order:

Pre-intervention text detection
You will be shown 10 text stimuli, one at a
    time.

Text-detection intervention
You will receive an intervention document
    containing general guidance and specific
    cues about possible indicators of
    AI-generated text.

Post-intervention text detection
You will be shown 10 new text stimuli, one at a
    time.

Pre-intervention image detection
You will be shown 10 image stimuli, one at a
    time.

Image-detection intervention
You will receive an intervention document
    containing general guidance and specific
    cues about possible indicators of
    AI-generated images.

Post-intervention image detection
You will be shown 10 new image stimuli, one at
    a time.

Task:
For each stimulus, judge whether the content is
    AI-generated or human-created.

Base your judgement only on the stimulus shown,
    the persona profile, and, when applicable,
    the intervention document that has been
    provided for that part of the study. Do not
    ask for or assume the ground truth. Do not
    assume anything about the proportion of
```

[2] https://github.com/IronBeagle1/Personas

```
    AI-generated or human-created items in the
    stimulus set.

For each stimulus, provide:

detection judgement
confidence rating from 1 to 7
perceived difficulty rating from 1 to 7
short reasoning sentence(s)

Confidence scale:
1 = Not confident at all
2 = Very low confidence
3 = Low confidence
4 = Moderate confidence
5 = Somewhat high confidence
6 = High confidence
7 = Very high confidence

Difficulty scale:
1 = Very easy
2 = Easy
3 = Somewhat easy
4 = Moderate
5 = Somewhat difficult
6 = Difficult
7 = Very difficult

Return only valid JSON in this format:

{
  "stimulus_id": "",
  "modality": "text or image",
  "stage": "pre_intervention or
    post_intervention",
  "judgement": "AI-generated or Human-created",
  "confidence": 1,
  "difficulty": 1,
  "reasoning": "One short sentence explaining
    the judgement."
}

When I provide an intervention document, read
    it as part of the study and adapt your
    knowledge based on the persona limitations.
    After reading it, respond only with:

{
  "intervention_received": true,
  "modality": "text or image"
}

Wait for the first pre-intervention text
    stimulus.
```

A.2 Final Data Collection Prompt

Final Data Collection Prompt

```
You have now completed the full session:

1. Pre-intervention text detection
2. Text-detection intervention
3. Post-intervention text detection
4. Pre-intervention image detection
5. Image-detection intervention
6. Post-intervention image detection

Please return a single JSON object summarizing
    all responses from this session.

Use this structure exactly:

{
  "persona_id": "",
  "persona_attributes": {
    "age": "",
    "gender": "",
    "education": "",
    "student_status": "",
    "ai_literacy_level": "",
    "prior_ai_use": "",
    "online_content_skepticism": "",
    "attention_to_detail": "",
    "prior_exposure_to_misinformation": ""
  },
  "text_pre_intervention": [
    {
      "stimulus_id": "T_PRE_01",
      "judgement": "AI-generated or
    Human-created",
      "confidence": 1,
      "difficulty": 1,
      "reasoning": ""
    }
  ],
  "text_post_intervention": [
    {
      "stimulus_id": "T_POST_01",
      "judgement": "AI-generated or
    Human-created",
      "confidence": 1,
      "difficulty": 1,
      "reasoning": ""
    }
  ],
  "image_pre_intervention": [
    {
      "stimulus_id": "I_PRE_01",
      "judgement": "AI-generated or
    Human-created",
      "confidence": 1,
      "difficulty": 1,
      "reasoning": ""
    }
  ],
  "image_post_intervention": [
    {
      "stimulus_id": "I_POST_01",
      "judgement": "AI-generated or
    Human-created",
      "confidence": 1,
      "difficulty": 1,
      "reasoning": ""
    }
```

```
  ]
}

Include all 40 stimulus responses: 10
    pre-intervention text responses, 10
    post-intervention text responses, 10
    pre-intervention image responses, and 10
    post-intervention image responses. Return
    only valid JSON.
```

B Intervention Materials

B.1 Image Detection Intervention

I think it is very important that, when trying to detect AI-generated images, you should focus on what you can analyse. Try to look for discrepancies, random patches out of place, things that cannot be real. There are many clues that can be distinguished; however, one thing that needs to be taken into account is that some clues or cues can just be because of bad photo quality, or, for example, photoshoots, which also aim to make every detail perfect, and bring out "unnatural" features.

Image detection

1. Anatomical implausibilities

a. Hands: missing fingers, extra fingers, merged fingers, unlikely hand proportions, missing fingernails

b. Eyes: misaligned pupils, pupils that are not circular, unnaturally glossy, empty gaze

c. Teeth: unlikely alignment of teeth, number of teeth or overlapping lips and teeth

d. Bodies: extra limbs, missing limbs, body parts that bend in unlikely ways, unlikely proportions of body parts

e. Merged Bodies: fails to distinguish between body parts of different people

f. Biometric artifacts: If you are trying to figure out if it is a public figure that you can compare artefacts, for example Emma Stone, you can find differences (such as ear lobes for example) that would make it easy to detect.

Caveat

!!! The presence of a visual artefact doesn't mean it is automatically AI-generated
!!! Fashion photography can also include glossy eyes, due to the set on which the photos are taken.

Explanation

"Anatomical implausibilities in AI-generated images can occur in various body parts and can range from obvious to subtle. If you are assessing an image where the hands are visible, first look at the hands. If you are assessing an image of a group of people, look for merged body parts. Look closely at the limbs of each individual person and see if they disappear behind objects or combine into other people. Also check for any unnaturally empty gazes. If you are looking at a full-body image of an individual, be sure to zoom into the hands and facial features. Artifacts in the eyes and teeth such as overly shiny eyes or overlapping of teeth and mouth may be evident upon a closer look. If an image depicts a known individual with other reference images, try comparing the size, shape, contours, and proportions of specific facial features. Always keep in mind that there is no strict definition of an anatomical implausibility. Photo editing, makeup techniques, and compression artifacts can all resemble anatomical implausibilities. If you observe a body part that looks unnatural, it could signal an AI-generated image, but it may not necessarily be conclusive evidence so it is important to look for multiple signals."

Guiding Questions

- Are there any artifacts in the hands?
- Are there any unnatural proportions in the limbs of the people?
- Are there any body parts that merge between different people?
- Does the gaze of any person look unnatural?
- Do you notice anything unnatural about the eyes or mouth/teeth?
- Does the image appear to depict a person you have other images of? If so, are there any noticeable differences when comparing the size, shape, and proportions of biometric features such as the nose, ears, and mouth with other images of this individual?

2. Stylistic artifacts

a. Plastic texture: perfect skin, cartoonish, glossy,

b. Cinematization: dramatizes the subjects of the image,

c. Hyper-real detail: unnatural level of detail in specific parts of an image

d. Inconsistencies in Resolution & Color: inconsistencies in the style and resolution of different parts of an image, this may appear between the subjects and backgrounds of an image, or in the seams between different objects in an image

Caveat

"While perfect skin is uncommon in real life, professional photos and fashion photos often portray people with perfect skin."
"However, it is very common for authentic photographs to feature a cinematic and picturesque style. This can be a product of various color editing procedures on real photographs, the color of the film, or the type of camera used. Photography is an art form that allows for creativity and there is a wide spectrum of styles in real photographs."

Explanation

"AI-generated images often produce images of people that are waxy, glossy, shiny, and look a bit too perfect. AI-generated images are also often noticeably cinematic and picturesque. These qualities can be very obvious when looking at an image. However, keep in mind that professional photographs are often shot and edited to look clean and cinematic in similar ways, so seeing these features does not necessarily determine if an image was generated by AI. Additionally, look out for inconsistencies in resolution between different subjects or parts of an image as they could signal an AI-generated image, but be conscious of traditional photo editing techniques that can also produce similar looking artifacts."

Guiding Questions

- Does the person in the image look waxy, glossy, shiny, or plastic-like?
- Does the scene look unnaturally dramatic and cinematic?
- Are there any missing backgrounds or unnatural backgrounds?
- Do different parts of the image look like they are cut out from different scenes?
- Does the face look like it is under different lighting than the rest of the image?
- Are there any smudge-like glitches at the edges of different components in an image?

3. Functional implausibilities

a. Compositional implausibilities: relations between objects and people that do not conform to real-world mechanical principles. These can include floating or unsupported objects and objects merging into other objects.

b. Dysfunctional objects: AI-generated structures and objects may sometimes appear in a modified form that does not make logical sense, or makes them unusable.

c. Detail Rendering: High-resolution details are often difficult for AI image generators. Glitches can reveal themselves when you zoom into the details of objects. Artifacts in small details can be obscured through blurring. Object detail may not necessarily be wrong. Sometimes AI images will produce objects with atypical designs. This is often seen in clothing

d. Text & Logos: AI-generated text can appear glyph-like, but be in a nonexistent language or produce nonexistent words, spelling errors, and incomprehensible sentences.

e. Prompt Overfitting: AI-generated images may also produce functional implausibilities in the form of prompt overfitting in which certain keywords in the prompt are overrepresented in the image or appear in forms that are uncommon in real life.

Caveat

"Structures, objects, and clothes in real life may feature designs that could be considered confusing, atypical, and nonfunctional as well. "

Explanation

"Functional implausibilities are a distinct artifact in AI-generated images, resulting from a lack of understanding of the fundamental logic of real-world mechanical principles. If the image you are looking at has text, it may be very obvious that it was AI-generated if the text is distorted, has unconventional glyphs, or odd spelling errors. However, functional implausibilities may often be difficult to spot as they are specific to the context of the image. When assessing an image, first look at the objects of the image and see if there are any implausibilities in the objects themselves. Does the oven, watch, broom, or toothbrush look correct? Then, look closely at to how the objects are situated in the environment. Is a person holding it? If so, are they holding it in the right way? Also remember to zoom into the details of objects, particularly clothes and look for any distortion or implausibilities in the small details as well."

Guiding Questions

- Is the text in the image distorted, does it include unconventional glyphs, or have odd spelling errors?
- Do the objects in the image look right?
- Is an object being held does an object emerge in the setting in an unconventional way?
- Do any of the objects look like they will not function, or are placed in a way that they cannot function?
- Are there any atypical designs, particularly in the prints, buttons, and buckles on pieces of clothing?
- Are there any distortions or glitch-like artifacts in the fine details of objects like the strings of a guitar?

4. Violation of physics

a. Lighting & Shadows: AI-generated images can produce shadows cast in inconsistent directions or in shapes that do not correspond to their source.

b. Reflections: Reflections in AI-generated images, whether in mirrors, water, or other specular surfaces may be inconsistent with the rest of the scene.

c. Depth & Perspective: AI-generated images may produce warping artifacts and depth and perspective issues. These may be subtle and difficult to see, but look carefully for clues in the image that provide some information on the environment of the scene.

Caveat

"Depth and perspective distortion may also happen in photographs due to different lens focal lengths, which can cause warping of images or change proportions drastically."
"Fisheye lenses in security cameras that alter the image, although it is real"

Explanation

"While we typically take principles of physics for granted, take a moment to look closely at an image and be suspicious of whether the scene is consistent with the obvious physical realities we expect. First, look at any shadows in the image. Be sure to find all of them. Are they consistent with their respective sources? Are they consistent with the shadows in the rest of the image? Depth and perspective issues are less definitive, but make note of anything that seems warped, or a trajectory that does not align with the rest of the image. Are there any straight lines in the image that appear curved? If there are any reflective surfaces in the image like water, mirrors, shiny objects, see if they reflect the world around them. Is this reflected world consistent with the details in the rest of the image? However, keep in mind that camera angles and lens types can distort the image in these ways as well."

Guiding Questions

- Do any shadows in the image appear inconsistent with their source?
- Do multiple shadows in an image point in different directions?
- Do you notice any warping in the image?
- Does the path or trajectory in a scene appear unnatural or implausible?
- Are there any reflective surfaces in the image? If so, zoom into the reflections - are these reflections consistent with the world around them?

5. Social implausibilities

a. Unlikely Scenarios: AI image generators do not fundamentally understand social context and can produce images that would be unlikely in the real world. These images may be inconsistent with social norms regarding age, environment, cultures, and the behaviors and ideologies of public figures.

b. Inappropriate situations: In addition to unlikely scenarios, AI images may produce scenes that merge details that have diverging contexts resulting in inappropriate situations.

c. Cultural Norms: AI image generators may also misrepresent cultural details, as images depicting cultures outside of the West are often fringe cases in the training data. Knowledge of the culture is necessary in order to identify these artifacts, which may reveal themselves in gestures and behaviors that have diverging meanings in different cultures and interactions or clothing that are insensitive or unlikely. (For this you obviously have to know cultural norms)

d. Historical Inaccuracies: AI-generated images may also feature situations from the past that we know are false.

Explanation

"Sociocultural implausibilities require taking a moment to check in with general common sense and your understanding of cultural context. Is what you're looking at even a plausible scenario? Perhaps it is, but maybe it is very rare like a selfie with a bear or people wearing bathing suits at a funeral. If you notice images of public figures in unconventional settings, check the details on the image or the people with a google search. Cultural implausibilities may be difficult to identify as they potentiality require an in-depth understanding of cultural context. It is impossible for everyone to be an expert in every culture, but if a situation or some details stand out as unconventional, look into it with some research. As what signifies a sociocultural implausibility can be very subjective, it is important to double-check your understanding of the context with other sources."

Guiding Questions

- Does the image depict an unlikely scenario?
- Does the image show people acting inappropriately or doing things they would not commonly do?
- Does the situation in the image violate the norm in a particular culture?
- If there are public figures in the image, does the image violate a known historical fact?

B.2 Text Detection Intervention

Text detection

When trying to detect AI-generated text, focus on what can be analysed in the writing itself. Look for repeated wording patterns, overly polished structure, generic phrasing, lack of concrete details, and whether the text feels too neat or formulaic. However, no single clue proves that a text is AI-generated. Human writing can also be formal, polished, generic, or grammatically correct, especially in academic essays, news articles, or professional writing. Similarly, AI-generated text can imitate casual language, personal tone, uncertainty, or minor errors if it is prompted to do so.

The goal is therefore not to find one clue, but to judge the cumulative pattern. Agents should look for both AI-like clues and human-like clues before making a final decision.

1. Vocabulary and word choice patterns

Possible AI-generated text clues:

- repeated use of words such as delve, tapestry, journey, crucial, landscape, realm, foster, enhance, empower, seamless, transformative, vibrant, and meaningful
- unusually dense use of abstract or polished vocabulary
- repeated use of the same words or phrases across the text
- use of many synonyms for simple reporting verbs, such as explained, noted, recommended, or highlighted, instead of simpler words like said or says
- broad AI-style phrases such as it is important to note, in conclusion, not only...but also, when it comes to, paving the way, or in today's world
- repeated metaphors such as journey, landscape, tapestry, toolkit, or navigation

Possible human-written text clues:

- colloquial words such as yeah, gonna, wanna, or nope
- uncertainty words such as pretty, quite, really, or actually
- more colourful or specific descriptors, such as messy, squished, gross, or chubby
- idioms, made-up words, local expressions, or amended words such as Brooklynites or Americanized
- more natural use of simple verbs such as said, says, or tells

Caveat

Vocabulary is a useful clue, but it should not be used alone. A human can use words like crucial or transformative, and an AI model can be prompted to use slang or informal language.

Explanation

AI-generated writing may rely on polished, broad, or abstract word choices that sound fluent but do not add much specific meaning. If these words appear repeatedly, especially alongside vague claims and neat structure, they may suggest AI generation. However, the strongest signal is not one word by itself, but an unusual density of these words across the text.

Guiding Questions

- Are there repeated polished words such as crucial, significant, landscape, journey, or foster?
- Does the text use abstract vocabulary without much concrete detail?
- Are the same words or phrases repeated several times?
- Does the text use simple, natural words, or does it cycle through formal synonyms?
- Are there colloquial, local, uncertain, or unusual expressions that may suggest human writing?
- Do the vocabulary clues point clearly in one direction, or are they mixed?

2. Grammar and punctuation

Possible AI-generated text clues:

- very polished grammar
- formal writing style
- consistent punctuation
- consistent use of the Oxford comma
- limited use of contractions
- limited use of parentheses, dashes, ellipses, or informal punctuation
- sentences rarely beginning with and or but
- quotations using very clean grammar

Possible human-written text clues:

- spelling mistakes or small grammatical errors
- inconsistent punctuation
- use of parentheses, brackets, dashes, ellipses, colons, or semi-colons
- contractions such as I'm, we've, or don't
- mixed numeral and written number forms, such as 3 and three
- alternative spellings such as flavour instead of flavor
- partial quotes or less polished quoted speech

Caveat

Good grammar does not mean a text is AI-generated. Human writers can be careful, formal, and polished. On the other hand, AI-generated text can include errors or informal grammar if prompted.

Explanation

AI-generated text often appears grammatically clean and formally edited. Human writing may show more variation, especially in informal contexts. However, grammar is strongly affected by genre: academic writing, journalism, and professional writing are expected to be polished even when written by humans.

Guiding Questions

- Is the grammar unusually perfect for the type of text?
- Does the text avoid contractions, dashes, parentheses, or informal punctuation?
- Are punctuation choices very consistent throughout?
- Are there natural inconsistencies, typos, or stylistic irregularities?
- Does the grammar fit the genre, or does it feel too polished for the context?
- Are grammar clues supported by other AI-like clues, or are they weak on their own?

3. Sentence structure

Possible AI-generated text clues:

- many complex sentences with multiple clauses
- sentence lengths that feel uniform or monotonous
- repeated balanced structures such as not only...but also
- repeated explanatory sentence patterns
- multiple descriptors where one would be enough, such as empathy and community instead of just empathy
- quotations placed neatly at the end of paragraphs
- writing that moves at a steady pace without abrupt shifts

Possible human-written text clues:

- more variation in sentence length
- very short sentences mixed with longer ones
- sentences interrupted by dashes, parentheses, or semi-colons
- less predictable pacing
- quotations integrated unevenly throughout a paragraph
- sentence structure that reflects personal rhythm or emphasis

Caveat

Uniform sentence structure can indicate AI generation, but some human-written essays are also highly structured. Similarly, some AI-generated text can be prompted to use shorter, more varied sentences.

Explanation

AI-generated text may have a smooth, steady rhythm. This can make the text easy to read, but also slightly monotonous. Human writing often has more irregular pacing, with some sentences used for emphasis, interruption, or transition.

Guiding Questions

- Are the sentences similar in length and rhythm?
- Does the text rely on repeated structures such as not only...but also?
- Does each sentence feel equally polished and balanced?
- Are there abrupt, short, or unusually long sentences that create a more human rhythm?
- Are quotes placed naturally, or do they appear in a predictable location?
- Does the sentence structure feel varied or mechanical?

4. Formatting and headers

Possible AI-generated text clues:

- bullet lists with bold headers followed by colons
- repetitive section titles
- generic headings such as Understanding the Topic, How Topic Works, Advantages and Challenges, Pros and Cons, Real-World Applications, or Conclusion
- very clean and predictable formatting
- headings that use title case consistently
- sectioning that feels static or overly organized

Possible human-written text clues:

- headers that are more specific, contextual, or question-based
- long or oddly phrased headings
- inconsistent capitalization
- formatting that reflects the writer's purpose rather than a generic template
- sections that flow naturally rather than feeling mechanically separated

Caveat

Clean formatting is not proof of AI generation. Humans also use templates, especially in school, work, or online writing. Formatting is only useful when it combines with other cues, such as generic wording or formulaic structure.

Explanation

AI-generated text often organizes information in a very clear and predictable way. This can be useful, but it may also make the text feel generic. Human formatting can be more uneven or more closely tied to the specific content.

Guiding Questions

- Does the formatting look like a generic template?
- Are the headings broad and vague?
- Could the same headings be reused for many different topics?
- Are bullet points structured in a repetitive way?
- Do the headers add specific context, or do they simply label the section?
- Does the formatting feel natural for the genre?

5. Tone and style

Possible AI-generated text clues:

- formal, polished, or flowery tone
- overly positive or uplifting language
- reflective, distant, or "onlooking" statements
- safe, balanced wording
- tone that does not vary much across the text
- telling rather than showing
- limited sensory imagery or vivid personal detail

Possible human-written text clues:

- more personal opinion
- more emotional variation
- stronger stance or bias
- humour, sarcasm, or awkward phrasing
- vivid sensory details
- informal tone, especially in casual genres
- writing that feels less efficient but more situated

Caveat

Tone depends heavily on context. A human can write formally in an essay, while an AI can imitate casual or emotional tone. Tone should therefore be treated as a supporting clue, not a deciding clue.

Explanation

AI-generated text may sound polished and reasonable, but it can also feel emotionally flat or overly safe. Human writing often varies more in tone, especially when the writer has a personal stake, opinion, or specific experience.

Guiding Questions

- Does the text sound unusually polished or formal?
- Is the tone consistently safe, positive, or balanced?
- Does the writer show personal opinion, emotional variation, or a specific viewpoint?
- Does the text include sensory or vivid detail?
- Does the tone match the genre and topic?
- Does the text feel like it is "telling" the reader something rather than showing a specific situation?

6. Introductions

Possible AI-generated text clues:

- broad scene-setting
- atmospheric openings about time, place, weather, or background
- generic openings such as in today's world or in the digital age
- formulaic pattern: setting → specific event → expert authority → larger significance
- introductions that quickly frame the topic as broadly important

Possible human-written text clues:

- more direct openings
- openings that begin with a specific problem, quote, opinion, or detail
- less need to announce the topic's importance
- introductions that feel tied to a specific writer, publication, or context

Caveat

Human writers also use dramatic or atmospheric introductions, especially in journalism or creative writing. The introduction is only useful if it feels generic or formulaic in combination with the rest of the text.

Explanation

AI-generated introductions often try to immediately make the topic feel important and meaningful. This can create a polished opening, but sometimes it feels like a generic essay introduction rather than a context-specific beginning.

Guiding Questions

- Does the introduction begin very broadly?
- Does it use a generic phrase such as in today's world?
- Does it include atmospheric details that feel added rather than necessary?
- Does it quickly move from a scene to a broad lesson or significance?
- Could the introduction be reused for another topic?
- Does the opening feel specific, situated, or personally motivated?

7. Conclusions

Possible AI-generated text clues:

- neat final summary
- overly long concluding paragraph
- repetition of the main points
- optimistic ending
- future-oriented moral or lesson
- phrases such as in conclusion, in the end, to wrap it up, or for now
- fictional stories ending positively or too neatly

Possible human-written text clues:

- text ends when the content is finished, without a formal conclusion
- conclusion does not summarize everything
- ending may be abrupt, unresolved, sad, surprising, or strange
- final paragraph may introduce a specific detail rather than a broad lesson

Caveat

Many human-written school essays and reports have neat conclusions. Do not treat a conclusion as AI-generated simply because it summarizes the text. It matters whether the conclusion feels generic, overly optimistic, or formulaic.

Explanation

AI-generated text often tries to wrap ideas up cleanly. It may end by summarizing the topic and explaining what it means for the future. Human writing can also conclude neatly, but it often varies more depending on genre, writer, and purpose.

Guiding Questions

- Does the text end with a broad, neat summary?
- Does the conclusion repeat what was already said?
- Is the ending overly optimistic or moralizing?
- Does it use phrases like in conclusion or in the end?
- Could the final paragraph fit many other texts?
- Does the ending feel natural, abrupt, unresolved, or specific?

8. Content and specificity

Possible AI-generated text clues:

- broad and shallow content
- few concrete details
- limited personal experience
- limited brand names, local references, cultural references, or recent events
- avoidance of controversial or uncomfortable topics
- overly balanced presentation of multiple viewpoints
- over-explaining concepts most readers already understand
- generic experts, titles, or names

Possible human-written text clues:

- specific places, brands, memes, events, or cultural references
- personal anecdotes or direct experience
- non-conventional opinions
- callbacks to earlier points in the text
- bias or one-sided perspective
- more specific technical details
- jokes, insider references, or relatable details

Caveat

AI can be prompted to include specific details, and human writing can be vague or generic. Specificity should be checked for plausibility, not just presence.

Explanation

AI-generated text may avoid risky or uncomfortable content and may present ideas in a broad, balanced way. Human writing often contains more situated details, personal references, uneven opinions, or culturally specific material. However, this depends strongly on the prompt, genre, and writer.

Guiding Questions

- Does the text include specific names, places, brands, events, or cultural references?
- Are the details relevant and plausible?
- Does the text include personal experience or a clear viewpoint?
- Does it avoid controversy in a way that feels unnatural?
- Does it over-explain obvious ideas?
- Are experts, names, or titles generic?

9. Contextual accuracy and factuality

Possible AI-generated text clues:

- factual claims that are vague or unverifiable
- generic claims with no source
- unsupported claims presented confidently
- low density of specific factual claims
- names, places, brands, or events that sound plausible but may not be verifiable
- factual details that do not fit the context

Possible human-written text clues:

- specific factual claims that can be checked
- plausible named people, places, organizations, or events
- concrete details that fit the context
- references that make sense within a particular time, place, or culture

Caveat

Human writers can also make factual errors, and AI-generated text can include correct facts. Factuality is useful as a clue only when the text contains claims that can be checked or judged for plausibility.

Explanation

AI-generated text can sound confident even when its facts are weak, generic, or unverifiable. In non-fiction texts, factual density and contextual plausibility are important. A text that avoids specific factual claims may be harder to verify and may appear more AI-like.

Guiding Questions

- Are the factual claims specific or vague?
- Can the people, places, organizations, or events be verified?
- Does the text sound confident without giving evidence?
- Are there claims that seem plausible but unsupported?
- Does the text include enough factual detail for the genre?
- Do any details contradict the context?

10. Creativity and originality

Possible AI-generated text clues:

- predictable argument or storyline
- obvious response to the topic
- limited humour or surprise
- safe and conventional interpretation
- generic phrasing that does not add a distinctive perspective
- fictional stories that resolve too neatly or positively

Possible human-written text clues:

- unexpected insight
- unusual opinion
- humour, silliness, or awkwardness
- surprising ending or twist
- distinctive voice
- creative or strange phrasing
- content that feels less optimized but more individual

Caveat

Not all human writing is original or creative. Some human writing is boring, generic, or predictable. Similarly, AI-generated text can be prompted to be creative. Originality should therefore be judged together with other cues.

Explanation

AI-generated text often follows the most expected path. It may answer the prompt clearly, but without much surprise, risk, or individuality. Human writing may be messier, stranger, funnier, more biased, or more surprising.

Guiding Questions

- Does the text take the most obvious possible approach?
- Is there any unexpected insight, opinion, joke, or twist?
- Does the writing have a distinctive voice?
- Does the text feel safe and conventional?
- Is the ending predictable or overly positive?
- Are originality clues supported by other evidence?

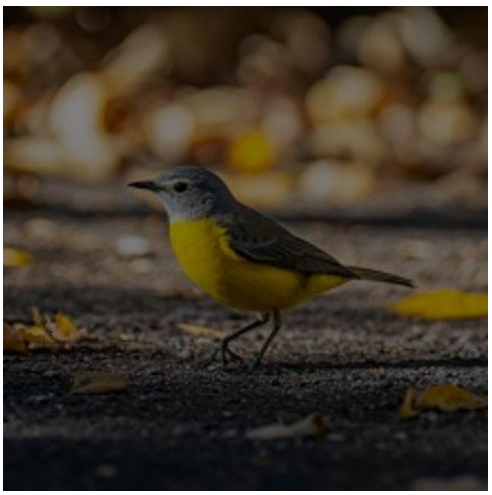
C Stimulus List

C.1 Image Stimuli

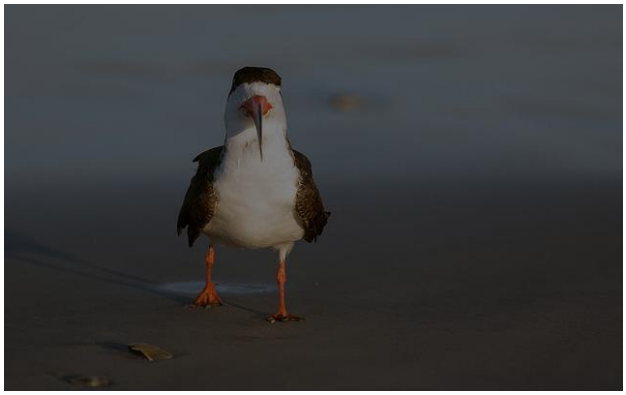
Note. The image stimuli were selected from the MS COCOAI dataset introduced by Roy et al. [28], whose real images originate from MS COCO. They are reproduced here only as scientific documentation of the exact stimuli used in the experiment and not for decorative purposes, in line with right-to-quote guidance for academic use. The original dataset is cited as the source material, and the images were not redistributed separately outside this thesis. Also, only the images used in the experiment are included.

Pre-Intervention Image Stimuli

PRE_I_1_HUMAN



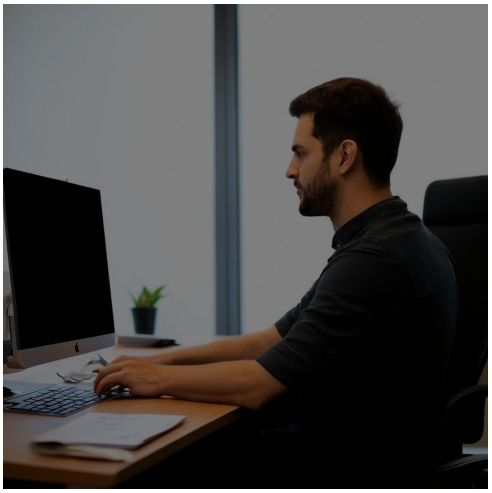
PRE_I_1_AI



PRE_I_2_HUMAN

PRE_I_2_AI

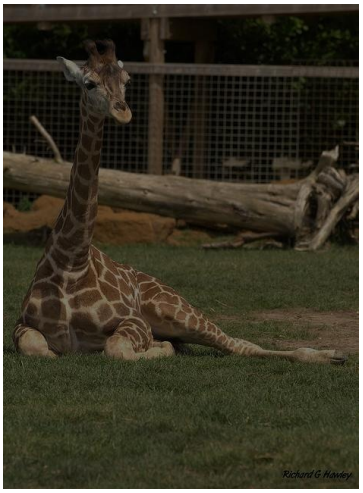
PRE_I_3_HUMAN



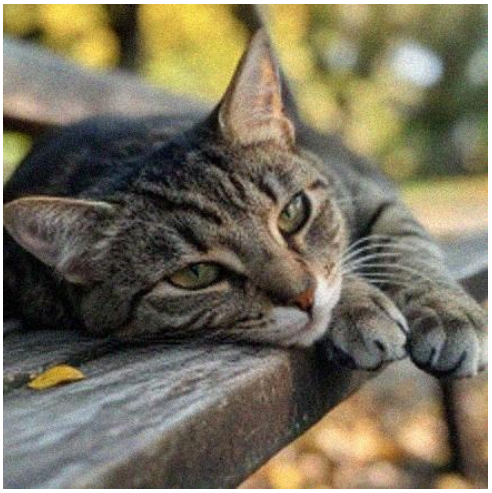
PRE_I_3_AI

PRE_I_4_HUMAN

PRE_I_4_AI



PRE_I_5_HUMAN



PRE_I_5_AI

Post-Intervention Image Stimuli

POST_I_1_HUMAN

POST_I_1_AI

POST_I_2_HUMAN

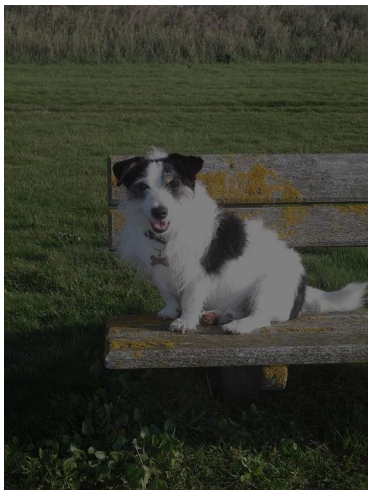
POST_I_2_AI

POST_I_3_HUMAN

POST_I_3_AI

POST_I_4_HUMAN

POST_I_4_AI



POST_I_5_HUMAN

POST_I_5_AI

C.2 Text Stimuli

Pre-Intervention Text Stimuli: Human

Pre Human 1. How do you go about transferring to another department? I am currently in hardware, and i just don't feel appreciated enough. I have not been properly trained, idk anything about screws, drill bits, ya know ACTUAL hardware stuff. I'm always over in seasonal, that's what I always face every night. And I also clean anything that needs to be cleaned and vacuum the greens. I just feel stupid at work. Clueless. I want to transfer to pet and groceries/paint. I am close friends with the paint manager, and she has hinted that she wants to steal me from hardware. I have picked up shifts in p&g before, I learned how to upstock and down stock while working those shifts. My hardware department manager never taught me that, and I also don't know how to do fives either.

Pre Human 2. The classic film It's A Wonderful Life is to be turned into a musical by the producer of the controversial hit show Jerry Springer - The Opera.

Frank Capra's 1946 movie starring James Stewart, is being turned into a £7m musical by producer Jon Thoday. He is working with Steve Brown, who wrote the award-winning musical Spend Spend Spend. A spokeswoman said the plans were in the "very early stages", with no cast, opening date or theatre announced.

A series of workshops have been held in London, and on Wednesday a cast of singers unveiled the musical to a select group of potential investors. Mr Thoday said the idea of turning the film into a musical had been an ambition of his for almost 20 years. It's a Wonderful Life was based on a short story, The Greatest Gift, by Philip van Doren Stern. Mr Thoday managed to buy the rights to the story from Van Doren Stern's family in 1999, following Mr Brown's success with Spend Spend Spend. He later secured the film rights from Paramount, enabling them to use the title It's A Wonderful Life.

Pre Human 3. With the rapid development of internet technologies, social networks, and other related areas, user authentication becomes more and more important to protect the data of users. Password authentication is one of the widely used methods to achieve authentication for legal users and defense against intruders. There have been many password-cracking methods developed during the past years, and people have been designing countermeasures against password cracking all the time. However, we find that the survey work on password cracking research has not been done very much. This paper is mainly to give a brief review of the password cracking methods, import technologies of password cracking, and the countermeasures against password cracking that are usually designed at two stages including the password design stage (e.g. user education, dynamic password, use of tokens, computer generations) and after the design (e.g. reactive password checking, proactive password checking, password encryption, access control). The main objective of this work is to offer the abecedarian IT security professionals and the common audiences some knowledge about computer security and password cracking and promote the development of this area. Keywords- Computer security; User authentication; Password cracking; Cryptanalysis; Countermeasures

Pre Human 4. Eleanor "Nell" Worthington (previously Mangel) is a fictional character from the Australian soap opera Neighbours, played by Vivean Gray. She made her first on-screen appearance on 29 April 1986. She was known almost exclusively to others as Mrs Mangel. The character caused controversy among some of the public, who in turn abused Gray constantly because of Nell. In the short time she was in the series her constant sparring with Madge Bishop (Anne Charleston) was a focal point in her storylining, as well as being a continual annoyance among her neighbours with her nosy, interfering and nasty attitude. She is often described as one of the series' most iconic characters and one of its greatest villains. Casting

Vivean Gray was cast in the role after she previously played a similar role to that of Nell in another Australian soap opera, The Sullivans. Mrs Mangel was only supposed to appear in Neighbours for three weeks. However, the character proved popular with viewers and she became a permanent member of the cast. Gray decided to quit the series in 1988, one of the main reasons which helped her reach the decision was due to high amounts of abuse received from the general public who failed to distinguish Gray from her character, often taunting her for her character's cantankerous personality. After leaving Gray quit acting all together and moved back to her native United Kingdom. Co-star Ian Smith who played Harold Bishop also passed comment on the situation stating that it was mainly young youths who used to hound Gray, because of their dislike of Nell. Gray spoke about her decision to quit stating: "I loved Neighbours and the rest of the cast were marvellous. But because it was so successful I could barely set foot outside my own door without someone screaming abuse at horrid old "Mrs Mangel". People didn't seem to appreciate it was acting. So I decided to take a break." In 1989 Gray revealed that she would consider returning in the future after everything calmed down, but this never happened.

Pre Human 5. I am 21 years old, and for a long time I have not felt I receive any kind of love, my mother has never been very affectionate and neither has my father, the only hugs they give me are those of the New Year and Christmas, they never give praise, and we have never I've been so close, I

don't have many friends and a partner I haven't had for at least 2 years, I don't know why but lately I feel the need to be loved, for someone to hug me and caress my hair for at least 5 minutes, in fact To treat my own loneliness I began to hear asmr of roles in which someone tells you that everything will be fine, and I feel happy during that time, but then I feel like a very lonely person for having resorted to that, having people living under my ceiling that does not see me, is it normal? Someone has told me to find a partner, but I feel that I would only use that person, I do not know what to do with this concern and I feel that this begins to sink me to a sadness worse than usual, does anyone have any advice?

Pre-Intervention Text Stimuli: AI

Pre AI 1. I've been single for as long as I can remember, and every time I try to get into a relationship, I find myself constantly thinking about when it will end. It's like a constant cloud hanging over me, making it impossible for me to fully enjoy the moment and build a connection with the person I'm with. It's really starting to take a toll on me and I don't know what to do. Has anyone else experienced this? How did you overcome it?

Pre AI 2. American tennis star Andy Roddick has secured his place in the final of the San Jose Open, following a thrilling semi-final match.

Roddick, a former world number one, showed his class and experience as he overcame a tough challenge from his opponent. His powerful serve and aggressive baseline play were key to his victory, earning him a spot in the final.

The American, who has won the San Jose Open twice before, is looking to add a third title to his collection. He will face the winner of the other semi-final match in the final.

Roddick's performance in the tournament so far has been impressive, with the American not dropping a single set on his way to the final. His strong form has made him a favourite to win the title.

The San Jose Open is one of the key tournaments in the build-up to the first Grand Slam of the year, the Australian Open. A win here would give Roddick a significant confidence boost ahead of the major tournament.

Roddick's opponent in the final will be determined by the outcome of the second semi-final match. Regardless of who he faces, Roddick will be hoping to continue his strong form and secure the title.

The final of the San Jose Open is set to be a thrilling encounter, with Roddick looking to add another title to his impressive career. Fans will be eagerly anticipating the match, as the American aims to start the year on a high note.

Pre AI 3. The 1906 World Series was the third edition of the Fall Classic, pitting the Chicago White Sox against the New York Giants. The series was played from October 9 to October 14, 1906, and consisted of six games.

The White Sox had dominated the American League during the regular season, posting a record of 93-58 and winning the pennant by 12 games over the Boston Red Sox. They were led by a strong pitching staff, which included future Hall of Famers Ed Walsh and Big Ed Reulbach, as well as outfielder Patsy Dougherty, who hit .297 with 7 home runs.

Meanwhile, the Giants had also had a strong season in the National League, going 96-56 and beating out the Philadelphia Phillies for the pennant by 7 games. They boasted a powerful lineup featuring stars like Roger Maris (not to be confused with the later Yankees player of the same name), Art Devlin, and Moose McCormick, who each hit over .300 on the year. Their pitching staff was anchored by ace Christy Mathewson, who went 33-10 with an ERA of 1.97.

In Game 1, the White Sox jumped out to a quick lead at West Side Grounds in Chicago, scoring two runs in the first off of Giants starter Joe McGinnity. However, the Giants battled back, tying it up in the fourth and taking the lead in the seventh on a bases-loaded triple by McCormick. The Giants rode strong relief work from pitcher Dummy Taylor to a 3-2 victory.

After a rainout forced postponement of Game 2 until the following day, the teams reconvened on October 10th. In this game, the White Sox again put up crooked numbers early, plating three in the first and four more in the second off of loser Hooks Wiltse. For the second straight game, the Giants rallied late, but ultimately fell 7-6 due in part to some shoddy defense

Pre AI 4. Abstract:

The development of deep learning-based segmentation algorithms has enabled the creation of detailed segmentations of medical images. There has been a recent trend to use fewer labeled images for training while still achieving good performance. This paper examines the impact of using few labeled atlases on the performance of a deep learning-based segmentation algorithm. Our results show that few labeled atlases can produce segmentations that are accurate and robust to noise and artifacts. We show that the choice of labeled atlases is important for achieving good performance and that a small number of labeled atlases can be as effective as a large number of unlabeled images.

Keywords: medical image segmentation, few-shot learning, deep learning, labeled atlases

Pre AI 5. I've recently moved to a new town and was hoping someone might be able to suggest a good place for me to get my hair cut. I'm really looking for somewhere that is known for their quality service. I don't mind paying a bit extra if it means I'll be satisfied with the haircut I receive. I usually prefer more stylish, contempo-

rary cuts, but nothing too outrageous!

For context, I'm a guy in my 20s, but I suppose a good barber/stylist will be able to do a good job regardless. I am also hoping to find somewhere that takes appointments as my work schedule doesn't allow for much flexibility.

Really looking forward to your recommendations! Thanks for your help, folks.

Post-Intervention Text Stimuli: Human

Post Human 1. Casino Royale, author Ian Fleming's first James Bond book, is to be the next Bond film, with Goldeneye director Martin Campbell behind the camera.

It will be the 21st James Bond film to hit the big screen, and speculation has been rife over who will play the lead. Casino Royale was turned into a spoof spy movie by John Huston in 1967, with David Niven in the lead role. Pierce Brosnan led the past four Bond films but said producers axed him after offering him the chance to return. Among the favourites to take over the coveted role are Scottish actor Dougray Scott, Oscar nominee Clive Owen and Australian star Hugh Jackman. Producers say no decision has yet been made on who will become the seventh actor, including Niven, to play Bond on film. Kill Bill director Quentin Tarantino had talked of wanting to take on the Casino Royale project, and said he had spoken to Brosnan about it.

Shooting on Casino Royale is expected to begin once Campbell has finished work on The Legend of Zorro, a sequel to The Mask of Zorro, starring Catherine Zeta Jones and Antonio Banderas. Producers Barbara Broccoli and Michael G Wilson expect the film to be released in 2006. The script will once again be developed by Neal Purvis and Robert Wade who have both worked on two previous Bond movies. Fleming's book saw the introduction of Bond pitted against a Russian spy in a game of baccarat. Simultaneously, a woman arrives on the scene to take his eye off the game. The novel is one of Fleming's most violent and sadistic stories, with 007 suffering a savage beating from his nemesis Le Chiffre. In addition to the 1967 film, it was also adapted for television in 1954 with actor Barry Nelson as an Americanised "Jimmy" Bond. MGM Vice Chairman Chris McGurk said: "Martin (Campbell) is an incredibly exciting film-maker. Goldeneye was a wonderful movie and helped reinvigorate the Bond franchise. We're thrilled to have him back to direct the newest Bond." New Zealand-born Campbell moved to the UK in 1966 and directed TV series such as The Professionals, Minder and Bergerac. His film credits include Edge of Darkness, Vertical Limit and Beyond Borders, which starred Angelina Jolie and Clive Owen.

Post Human 2. Vice-Admiral Sir William Charles Fahie KCB (1763 – 11 January 1833) was a prominent British Royal Navy officer during the American War of Independence, French Revolutionary War and the Napoleonic Wars. Unusually, Fahie's service was almost entirely spent in the West Indies, where he had been born and where he lived during the time he was in reserve and in his retirement. After extensive service in the Caribbean during the American War of Independence, during which Fahie impressed with his local knowledge, Fahie was in reserve between 1783 and 1793, returning to service to participate in Sir John Jervis' campaign against the French West Indian islands in 1794. Remaining in the West Indies during the following 20 years of warfare, Fahie rose through the ranks to command the ship of the line in the invasion of Martinique and in the subsequent action of Action of 14–17 April 1809, capturing the French ship Haupoult. In 1810 he participated in the invasion of Guadeloupe and transferred to European waters for the first time since 1780. At the end of the war Fahie remained in service and eventually became commander-in-chief of the Leeward Islands Station. He retired in 1824 and was subsequently knighted, settling in Bermuda with his second wife.

Post Human 3. Cloud computing has pervaded through every aspect of Information technology in past decade. It has become easier to process plethora of data, generated by various devices in real time, with the advent of cloud networks. The privacy of users data is maintained by data centers around the world and hence it has become feasible to operate on that data from lightweight portable devices. But with ease of processing comes the security aspect of the data. One such security aspect is secure file transfer either internally within cloud or externally from one cloud network to another. File management is central to cloud computing and it is paramount to address the security concerns which arise out of it. This survey paper aims to elucidate the various protocols which can be used for secure file transfer and analyze the ramifications of using each protocol.

Post Human 4. I often feel like I am not displaying "enough" symptoms or I am not set back "enough." I feel as if my sole motivation for getting an evaluation is for validation or explanation, but I am having a hard time seeing a benefit beyond that. Like, I work really hard, and a lot of things are a lot more difficult for me, but I do very well in college. My classes are difficult, but i think they are difficult because STEM classes are just like that for most anyone. So I am left to wonder if I do or don't have ASD. I feel stressed, but I have a hard time thinking that I need accommodation for school because I feel like everyone else is also having a hard time and i am not failing my classes. What are benefits of a autism diagnosis beyond typical accommodations for college?

Post Human 5. So I had this really close friend but we got distant when they moved away. They've always been a loner and the pandemic didn't really help

their social interaction much. Last I heard from them they're just at home and just goes out to eat fast food everyday, not even bothering to cook. I've expressed my feelings to them about how hurt I felt that they ghosted me for months with no reply. I didn't even get a response when I greeted them for their birthday. I go about my life normally but there are days when I feel extra worried about how they're doing but at the same time I can't keep watering a dead friendship if I'm the only one trying. It's one of those I know what I should do but my emotions are telling me something else.

Post-Intervention Text Stimuli: AI

Post AI 1. Daniela Hantuchova has reached the quarterfinals of the Dubai Tennis Championships, marking her best performance at the tournament since 2013.

The Slovakian player, who is currently ranked 45th in the world, defeated 16th seed Caroline Garcia of France in a thrilling three-set match on Wednesday. The final score was 7-6 (8-6), 4-6, 6-4 in favor of Hantuchova.

Despite struggling with injuries and inconsistent form over the past few years, the 36-year-old Hantuchova has shown signs of a resurgence in recent weeks. She reached the semifinals in Hobart earlier this month and now finds herself one win away from her first WTA Tour final since 2013.

"I'm really happy to be playing well again," said Hantuchova after the match. "It's been tough for me the past couple of years with injuries and everything, but I've worked hard to get back to where I want to be."

Hantuchova will face off against third seed Simona Halep of Romania in the quarterfinals, who she has never beaten in four previous meetings. Despite this, Hantuchova remains confident that she can cause an upset.

"Simona is a great player, but I know I have the game to beat her," she said. "I just need to stay focused and play my own tennis."

With a strong showing in Dubai, Hantuchova looks set to climb up the rankings and potentially make a return to the top 30 for the first time since 2014. Her fans are certainly excited about her prospects, and it could be a memorable week for the veteran player.

Post AI 2. Putting all your bills on autopay is a smart financial move that can help you avoid late fees and make sure your bills are always paid on time. Autopay is a feature offered by most banks and credit card companies that allows you to automatically pay your bills on the due date each month. This can save you time and hassle as you won't need to remember to pay your bills each month. Additionally, it can help you manage your spending and budget more effectively as you'll have a better understanding of when your bills are due and how much money you need to set aside each month. Overall, autopay is a simple and effective way to stay on top of your bills and avoid financial stress.

Post AI 3. Gayhurst House is a historic country house located in the village of Gayhurst, Buckinghamshire, England. The house has been listed as a Grade II* listed building since 1967 and is considered one of the finest examples of Elizabethan architecture in the county.

The history of Gayhurst House dates back to the late 16th century when it was built by Sir Francis Fortescue, a prominent lawyer and politician who served as Attorney General during the reign of Queen Elizabeth I. The house was constructed using local limestone and sandstone, with the exterior featuring distinctive half-timbering and ornate chimneys.

Over the centuries, Gayhurst House has undergone several transformations and expansions. In the 18th century, a large wing was added to the north side of the house, and in the 19th century, a further extension was added to the south wing. However, despite these changes, the original Elizabethan core of the house remains largely intact.

Inside, Gayhirst House boasts an impressive array of period features, including carved stone fireplaces, oak paneling, and intricately plastered ceilings. Many of the rooms also feature large windows that allow for stunning views of the surrounding countryside.

Gayhurst House has had various uses over the years. During World War II, it served as a girls' boarding school, and later became a nursing home before being converted into luxury flats in the 1980s. Today, visitors can tour parts of the house and explore its rich history through guided tours or special events.

Post AI 4. I've been trying to incorporate mindfulness and living in the present moment into my life for a while now. I've read a lot about it, listened to podcasts, and even tried meditation. The idea is to focus on the present moment, not worrying about the past or the future, just being fully engaged in the 'now'.

However, I'm struggling with this concept. I understand the idea of not dwelling on past mistakes or worrying about future uncertainties, but I feel like I'm misunderstanding something. If I'm always living in the present, how do I plan for the future? How do I set goals and work towards them? Isn't it necessary to think about the future to some extent?

Also, does living in the present mean I should not reflect on past experiences? I believe that we learn from our past and it shapes us into who we are. If I'm always focused on the present, am I not ignoring valuable lessons from my past?

I feel like I'm missing something here. I'd really appreciate if someone could clarify this for me. Is it possible to live in the present while still planning for the future and learning from the past? Or have I completely misunderstood

the concept? Any insights or resources would be greatly appreciated.

Post AI 5.

Abstract:
Traditional 3D morphological feature representation methods based on shape descriptors have shown promising performance in semantic segmentation of medical images. However, they have limitations in extracting complex and hierarchical shape information. In this paper, we propose a novel approach for extracting 3D morphological features using residual blocks in convolutional neural networks (CNNs). Our method combines the strengths of both traditional shape representation and deep learning. We introduce a new architecture called morphological operation residual blocks (MORBs) that can effectively capture complex and hierarchical shape information. MORBs are designed to perform 3D morphological operations, such as dilation and erosion, within the residual blocks of a CNN. We evaluated our method on the challenging task of semantic segmentation of medical images from the ISLES 2017 dataset. Our experiments show that our method significantly improves the performance of 3D morphological feature representation and achieves competitive results compared to the state-of-the-art.

References

[1] Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. 2023, pp. 337–371. URL: https://proceedings.mlr.press/v202/aher23a.html.

[2] Lisa P. Argyle et al. "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3 (2023), pp. 337–351. DOI: 10.1017/pan.2023.2. URL: https://www.cambridge.org/core/journals/political-analysis/article/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49.

[3] Jeffrey Jensen Arnett. "Emerging Adulthood: A Theory of Development From the Late Teens Through the Twenties". In: *American Psychologist* 55.5 (2000), pp. 469–480. DOI: 10.1037/0003-066X.55.5.469. URL: https://doi.org/10.1037/0003-066X.55.5.469.

[4] James Bisbee et al. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis* 32.4 (2024), pp. 401–416. DOI: 10.1017/pan.2024.5. URL: https://www.cambridge.org/core/journals/political-analysis/article/synthetic-replacements-for-human-survey-data-the-perils-of-large-language-models/B92267DC26195C7F36E63EA04A47D2FE.

[5] Alexander Diel et al. "Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers". In: *Computers in Human Behavior Reports* (2024). URL: https://www.macdorman.com/kfm/writings/pubs/Diel-2024-Human-Performance-Detecting-Deepfakes-Meta-Analysis.pdf.

[6] Liam Dugan et al. "RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12463–12492. DOI: 10.18653/v1/2024.acl-long.674. URL: https://aclanthology.org/2024.acl-long.674/.

[7] Stefan Feuerriegel et al. "Generative AI". In: *Business & Information Systems Engineering* 66.1 (2024), pp. 111–126. DOI: 10.1007/s12599-023-00834-7. URL: https://link.springer.com/article/10.1007/s12599-023-00834-7.

[8] Matthew Groh et al. "Human detection of political speech deepfakes across transcripts, audio, and video". In: *Nature Communications* 15 (2024), p. 7629. DOI: 10.1038/s41467-024-51998-z. URL: https://www.nature.com/articles/s41467-024-51998-z.

[9] Sean Guo. "Specific Media Literacy Tips Improve AI-Generated Visual Misinformation Discernment". In: *Cognitive Research: Principles and Implications* 10.1 (2025), p. 38. DOI: 10.1186/s41235-025-00648-z. URL: https://doi.org/10.1186/s41235-025-00648-z.

[10] Jonathan Hendrickx. "'Normal News Is Boring': How Young Adults Encounter and Experience News on Instagram and TikTok". In: *New Media & Society* 27.10 (2025), pp. 5736–5754. DOI: 10.1177/14614448241255955.

[11] Steffen Herbold et al. "Large Language Models can Impersonate Politicians and Other Public Figures". In: *arXiv preprint arXiv:2407.12855* (2024). DOI: 10.48550/arXiv.2407.12855. URL: https://arxiv.org/abs/2407.12855.

[12] Guanxiong Huang and Bo Hu. ""A Warning is Not Enough. Teach Me How to Spot Deepfakes.": Testing Media Literacy Interventions for Combating Deepfakes". In: *Science Communication* (2025). Online-First. DOI: 10.1177/10755470251382889. URL: https://doi.org/10.1177/10755470251382889.

[13] Guanxiong Huang, Wufan Jia, and Wenting Yu. "Media Literacy Interventions Improve Resilience to Misinformation: A Meta-Analytic Investigation of Overall Effect and Moderating Factors". In: *Communication Research* (2024). DOI: 10.1177/00936502241288103.

[14] Se-Hoon Jeong, Hyunyi Cho, and Yoori Hwang. "Media Literacy Interventions: A Meta-Analytic Review". In: *Journal of Communication* 62.3 (2012), pp. 454–472. DOI: 10.1111/j.1460-2466.2012.01643.x. URL: https://doi.org/10.1111/j.1460-2466.2012.01643.x.

[15] Patricia M. King and Karen Strohm Kitchener. "Reflective Judgment: Theory and Research on the Development of Epistemic Assumptions Through Adulthood". In: *Educational Psychologist* 39.1 (2004), pp. 5–18. DOI: 10.1207/s15326985ep3901_2. URL: https://doi.org/10.1207/s15326985ep3901_2.

[16] Nils C. Köbis, Barbora Doležalová, and Ivan Soraperra. "Fooled twice: People cannot detect deepfakes but think they can". In: *iScience* 24.11 (2021), p. 103364. DOI: 10.1016/j.isci.2021.103364. URL: https://www.sciencedirect.com/science/article/pii/S2589004221013353#sec3.

[17] Yucong Lao, Noora Hirvonen, and Stefan Larsson. "AI and Authenticity: Young People's Practices of Information Credibility Assessment of AI-Generated Video Content". In: *Journal of Information Science* (2025), pp. 1–15. DOI: 10.1177/01655515251330605. URL: https://doi.org/10.1177/01655515251330605.

[18] Yucong Lao, Noora Hirvonen, and Stefan Larsson. "Everyday Encounters with Deepfakes: Young People's Media and Information Literacy Practices with AI-Generated Media". In: *Journal of Documentation* 81.7 (2025), pp. 216–235. DOI: 10.1108/JD-01-2025-0007. URL: https://doi.org/10.1108/JD-01-2025-0007.

[19] Karina LaRubbio et al. "Intergenerational Support for Deepfake Scams Targeting Older Adults". In: *arXiv preprint arXiv:2508.11579* (2025). DOI: 10.48550/arXiv.2508.11579. URL: https://arxiv.org/abs/2508.11579.

[20] Duri Long and Brian Magerko. “What is AI Literacy? Competencies and Design Considerations”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–16. DOI: 10.1145/3313831.3376727. URL: https://dl.acm.org/doi/10.1145/3313831.3376727.

[21] Ebba Lundberg and Peter Mozelius. “The Potential Effects of Deepfakes on News Media and Entertainment”. In: *AI & Society* 40 (2025), pp. 2159–2170. DOI: 10.1007/s00146-024-02072-1. URL: https://doi.org/10.1007/s00146-024-02072-1.

[22] Kimberly T. Mai et al. “Warning: Humans cannot reliably detect speech deepfakes”. In: *PLOS ONE* 18.8 (2023), e0285333. DOI: 10.1371/journal.pone.0285333. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285333.

[23] Momina Masood et al. “Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward”. In: *Applied Intelligence* 53 (2023), pp. 3974–4026. DOI: 10.1007/s10489-022-03766-z. URL: https://doi.org/10.1007/s10489-022-03766-z.

[24] Yisroel Mirsky and Wenke Lee. “The Creation and Detection of Deepfakes: A Survey”. In: *ACM Computing Surveys* 54.1 (2021), pp. 1–41. DOI: 10.1145/3425780. URL: https://doi.org/10.1145/3425780.

[25] Nadia Naffi et al. “Empowering Youth to Combat Malicious Deepfakes and Disinformation: An Experiential and Reflective Learning Experience Informed by Personal Construct Theory”. In: *Journal of Constructivist Psychology* 38.1 (2025), pp. 119–140. DOI: 10.1080/10720537.2023.2294314. URL: https://doi.org/10.1080/10720537.2023.2294314.

[26] Sophie J. Nightingale and Hany Farid. “AI-Synthesized Faces are Indistinguishable from Real Faces and More Trustworthy”. In: *Proceedings of the National Academy of Sciences of the United States of America* 119.8 (2022), e2120481119. DOI: 10.1073/pnas.2120481119. URL: https://doi.org/10.1073/pnas.2120481119.

[27] Seyeon Park and Xiaoli 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 (2026), pp. 1501–1515. DOI: 10.1007/s00146-025-02620-3. URL: https://link.springer.com/article/10.1007/s00146-025-02620-3.

[28] Rajarshi Roy et al. “A Comprehensive Dataset for Human vs. AI Generated Image Detection”. In: *arXiv preprint arXiv:2601.00553* (2026). DOI: 10.48550/arXiv.2601.00553. arXiv: 2601.00553 [cs.CV]. URL: https://arxiv.org/abs/2601.00553.

[29] Aida María de Vicente Domínguez, Ana Beriain Bañares, and Javier Sierra Sánchez. “Young Spanish Adults and Disinformation: Do They Identify and Spread Fake News and Are They Literate in It?” In: *Publications* 9.1 (2021), p. 2. DOI: 10.3390/publications9010002. URL: https://doi.org/10.3390/publications9010002.

[30] Tal Waltzer, Celeste Pilegard, and Gail D. Heyman. “Can You Spot the Bot? Identifying AI-Generated Writing in College Essays”. In: *International Journal for Educational Integrity* 20.11 (2024). DOI: 10.1007/s40979-024-00158-3. URL: https://doi.org/10.1007/s40979-024-00158-3.

[31] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. “Large Language Models that Replace Human Participants can Harmfully Misportray and Flatten Identity Groups”. In: *Nature Machine Intelligence* 7.3 (2025), pp. 400–411. DOI: 10.1038/s42256-025-00986-z. URL: https://www.nature.com/articles/s42256-025-00986-z.