REAL OR AI-GENERATED?

Testing AI and media literacy interventions with AI agents impersonating young adults

AUTHOR

Iustin-Nicolae Tudor
I.N.Tudor@student.tudelft.nl

COORDINATORS

Dr. Ujwal Gadiraju
Dr. Marije van Dalen
Shreyan Biswas
Esra de Groot

01. INTRODUCTION

Generative AI can produce realistic text and images, making human-created and AI-generated content harder to tell apart. This study uses AI agents impersonating young adults to explore whether AI and media literacy guidance can support detection.

02. RESEARCH FOCUS

- Intervention effect: Does guidance improve detection accuracy?
- Modality effect: Do results differ between text and image stimuli?

03. METHODOLOGY

Controlled simulation experiment: AI agents impersonating young adults complete detection tasks before and after an AI and media literacy intervention.

- **Pre-intervention:** identify AI-generated and human-created examples.
- **Intervention:** short guidance with detection tips.
- **Post-intervention:** classify new examples after the intervention.
- **Main measure:** change in detection accuracy

05. RESULTS

Detection accuracy improved after the intervention

OVERALL 85.75% → 94.25%

TEXT

	Pre-intervention	Post-intervention
Accuracy	94.5%	99.0%
Confidence	5.27	5.66
Difficulty	3.64	3.31

IMAGE

	Pre-intervention	Post-intervention
Accuracy	77.0%	89.5%
Confidence	5.15	5.45
Difficulty	3.75	3.52

Main pattern: Images showed the largest improvement, while text had the highest accuracy.
Remaining challenge: Most errors came from AI-generated content being judged as human-created.

04. EXPERIMENT DATA

- 20 AI-agent personas
- 800 total judgements
- Text + image stimuli
- Pre- and post-intervention design
- Measures: accuracy, confidence, difficulty

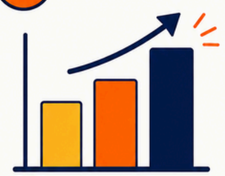
CONFIDENCE PATTERN

Correct vs incorrect judgments

1 Before the intervention

Correct: 5.26
Incorrect: 4.89
Gap: 0.37

2 After the intervention



Correct: 5.58
Incorrect: 5.17
Gap: 0.41

3 Main takeaway

Confidence increased, but incorrect judgments still received relatively high confidence ratings.

06. CONCLUSIONS & FUTURE WORK

- **Conclusions:** Cue-based guidance improved AI-generated content detection, especially for images. However, confidence did not clearly align with correctness, so it should be interpreted carefully.
- **Future work:** Test the intervention with real young adults, compare them with AI-agent personas, and evaluate the design across multiple AI models.