# Student Attitudes and Skills in a World of AI Hallucinations ⋆

Canh Thien, Dang[1], An, Nguyen[2]

[1]King's Business School, King's College London
[2]King's Business School, King's College London

**Abstract**

The rapid advancement of generative artificial intelligence (GenAI) has transformed educational and professional landscapes, offering new tools for learning and innovation. However, alongside its potential benefits, the phenomenon of AI hallucinations—false or fabricated information generated by these systems—has raised critical concerns. Misinformation created by GenAI poses a challenge to users' trust in AI and their willingness to adopt these tools effectively. In educational contexts, where students are increasingly exposed to AI technologies, understanding how these systems are perceived becomes crucial for shaping effective pedagogical strategies. Our study aims to make two contributions. First, we advocate for a pedagogical innovation in incorporating GenAI into assessment: training and evaluating students on evaluating information generated by GenAI, particularly on articulating and detecting AI hallucination. Second, we explore the interaction between students' ability to detect AI hallucinations and their resulting attitudes toward AI. By investigating the effects of exposing students to the prevalence of GenAI misinformation and their peers' detection capabilities, this research aims to uncover how such awareness influences trust, confidence, and the perceived usefulness of AI tools. The motivation stems from the need to address potential stereotypes and societal concerns about the reliability of AI technologies, which may deter students from embracing these tools in their learning environments. Through a randomised experimental design, this study seeks to isolate the causal mechanisms behind these dynamics, contributing valuable insights into fostering a balanced and informed approach to GenAI adoption in education.

**Keywords**

AI Hallucination, student skills, student attitudes

## 1. Introduction

With the release of Chat Generative Pre-Trained Transformer (ChatGPT) to the public via a user-friendly web interface in November 2022, Large Language Models (LLM) have emerged as a promising application to aid and augment work in a significant way, much more so in educational practices (Fütterer et al., 2023). Students can a benefit by conversing with and developing their subject understanding using the tools, and so quickly they did (The Times, 2023). The proliferation of available tools and potential uses poses challenges for educators and in many ways the society. One particular concern for educators is to embrace GenAI in creating effective assessments and maintaining academic integrity. A broader concern to the society is GenAI's tendency to generate false information that can often be masked under coherent and eloquent writing. If undetected, unverified, and unrectified, such false information can be inadvertently used or misused to various degrees of danger. In this paper, we propose the first experimentation to study whether and how students in a top UK business school can detect false information created by GenAI, which is often defined as AI hallucinations, in a high-stake assessment context. While we constrain our paper within the educational context, it is highly relevant to the emerging research on identifying the key traits and socioeconomic factors underlying news readers in recognising false information and fake news (Angelucci & Prat, 2024). Our setting presents a situation when readers (students) have abundant resources and training, as well as vested interest, to investigate and evaluate the information (AI-generated response to an assessment question). We aim to shed light on the extent to which economics and business-related courses educators can evaluate students' academic

*2nd International Workshop on AI in Society, Education and Educational Research (AISEER)*

⋆ **Preliminary results - expecting changes to the conceptual framework.**

*Corresponding author: canh.dang@kcl.ac.uk.

†These authors contributed equally.

performance considering the recent development in GenAI and proctored, in-person examination settings are to be avoided. Our evidence on students' ability to detect incorrect information beyond cohesive and well-prosed responses from GenAI contributes to the scholarship of teaching and learning on AI literacy in the education setting.

## 2. Literature Review

Existing research has introduced substantial opportunities within the educational landscape, benefiting both educators and learners in multifaceted ways. For educators, the integration of AI can significantly reduce the administrative burden by automating tasks such as report writing, lesson planning, content creation, and marking, as highlighted by Holstein et al. (2018). This automation not only frees up valuable time but also allows educators to focus more on direct student interaction and personalised teaching. Furthermore, AI can enhance the quality of educational materials, in a way that AI can introduce innovative ideas and approaches to teaching, ensuring that lessons remain engaging and effective. Additionally, AI's role in providing timely and detailed feedback through automated marking systems, as discussed by Hooda et al. (2022), reduces the marking workload for teachers and allows them to concentrate on more critical teaching tasks. Finally, AI supports teachers' continuous professional development by providing access to the latest pedagogical research and strategies. Nazaretsky et al. (2025) demonstrate how AI can analyse vast amounts of educational research and present actionable insights, helping educators stay current with the latest developments in teaching. Relying on personality measures, Stein et al. (2024) show that agreeableness and younger age predict more positive attitudes towards AI, whereas females are in general more cautious, so are those susceptible to conspiracy beliefs.

For students, AI offers personalised learning experiences by generating explanations of complex concepts and providing immediate support. This adaptive teaching assistance tailors learning materials to meet individual needs, offering advanced resources for those who require them. Additionally, the provision of one-to-one virtual tutoring through AI, explored by Chen & Shu (2023) and Liu et al. (2022), ensures personalised support by tailoring lessons to the unique learning pace and style of each student, thus addressing specific learning gaps and enhancing academic performance. Furthermore, AI creates immersive learning experiences through simulations and virtual environments, making learning more interactive and engaging (Liu et al., 2022). Enhancing accessibility and inclusion is another significant benefit of AI, particularly for foreign language students and students with special educational needs and disabilities (SEND). AI-assisted writing tools can assist foreign language students with writing and language tasks (Alharbi, 2023; Hwang et al., 2023), while Sharma & Dash (2023) highlight AI's ability to provide differentiated instruction tailored to students with specific needs. Additionally, AI supports employability and career development by offering personalised career advice and skills training (Mironko & Sutyniec, 2024; Ngotngamwong, 2020; Rožman et al., 2023). These benefits highlight the transformative potential of AI in education, enhancing both teaching and learning experiences significantly.

The attitudes towards the use of Generative AI (GenAI) in education are varied and reflect a spectrum of optimism and caution, influenced by geographical and contextual factors. In the UK, the general sentiment is broadly optimistic, with a small minority expressing unconditionally pessimistic views and doubting whether AI concerns can ever be fully mitigated (Bright et al., 2024). This contrast is mirrored in Norway, where there is a notable enthusiasm for GenAI among students, while professors adopt a slightly more cautious stance, indicating a potential generational gap in attitudes towards AI integration in education (Rasmussen & Karlsen, 2023). In Hong Kong, a survey of 399 undergraduate and postgraduate students from various disciplines revealed a generally positive attitude towards the use of GenAI in teaching and learning. However, these positive perceptions are tempered by concerns regarding accuracy, privacy, ethical issues, and the broader impact on personal development, career prospects, and societal values (Chan & Hu, 2023). This indicates a balanced view where the potential benefits of GenAI are acknowledged, but not without reservations about its implications. Contrastingly, in the African context, particularly within distance learning environments, lecturers

tend to be more pessimistic and cautious about the impact of tools like ChatGPT on academic integrity (Sevnarayan & Potter, 2024). In contrast, administrative staff and students exhibit a more transformative view, recognising GenAI's potential to enhance learning, albeit with the caveat that it requires careful management to avoid misuse (Sevnarayan & Potter, 2024). This divergence in attitudes suggests a divide between those directly responsible for maintaining academic standards and those focused on leveraging AI for educational innovation. In American universities, the trust in GenAI among students varies depending on their levels of confidence and motivation (Amoozadeh et al., 2024). This variability indicates that individual differences significantly influence how GenAI is perceived and utilised in educational contexts. Students with higher confidence and motivation levels are more likely to trust and effectively engage with GenAI tools, while others may remain sceptical or cautious.

These mixed attitudes towards GenAI in education highlight the complexities involved in its integration. While there is a broad recognition of its potential benefits, concerns about ethical implications, accuracy, privacy, and the impact on traditional educational values persist. This emphasises the need for a nuanced approach to GenAI implementation in education, one that carefully balances innovation with rigorous oversight and ethical considerations. The integration of AI in education has prompted concerns regarding over-reliance on GenAI tools, potentially hindering students' skill acquisition and impeding the quality of teaching. (Hyde et al., 2024) focused on the risk of excessive dependence on GenAI tools, which may limit students' ability to develop critical skills and knowledge acquisition.

## 3. What is AI Hallucination?

The concept of AI hallucinations, while still evolving, encompasses various manifestations arising from the use of large-language models (LLMs). Zhang et al., (2023) discuss three main types of AI hallucinations: (i) input-conflicting, (ii) context-conflicting, and (iii) fact-conflicting. Input-conflicting hallucinations occur when there is a contradiction between the content generated by LLMs and user inputs, such as when LLMs mistakenly perform tasks unrelated to user requests. Context-conflicting hallucinations arise from inconsistencies and irrelevant responses generated by LLMs, often due to their limitations in maintaining long-term coherence and handling complex instructions. An example of this is when LLMs display a lack of consistency in addressing individuals' names throughout a text. Fact-conflicting hallucinations involve fabricated or misleading information presented to users, contradicting established knowledge and potentially leading to misconceptions. Examples include incorrect answers that deviate from factual accuracy and established truths. In this paper, we focus on the third type of AI hallucinations: GenAI produces responses that are factually incorrect and inconsistent with the subject knowledge taught to students.

There are risks associated with AI hallucinations (Salvagno et al., 2023; Zhang et al., 2023). First, hallucinated responses are presented as ambiguous, which causes users difficulties in selecting the correct interpretation of the response (Zhang et al., 2023). This may decrease the usefulness of AI-assisted models and require effort from users to examine the possibilities. Another issue with AI-generated misinformation is that potential biases are inherent in the training data that only AI-literate users can detect (Sovrano et al., 2023). As such, AI hallucinations may exacerbate the existing stereotypes, discrimination, and societal biases of the training dataset into further reproduction of LLMs responses. Lastly, AI hallucinations, especially the context-conflicting type, often provide users with fragmented and inconsistent information, which can adversely affect online safety and public trust (Chen & Shu, 2023).

Strategies aimed at mitigating concerns surrounding AI in education can be broadly categorized into two approaches: intervention-based and a more cautious approach. Intervention-based strategies involve the implementation of policies and guidance to regulate the use of AI tools in educational settings. These policies aim to provide clear frameworks for the ethical and responsible use of AI, thus addressing concerns related to accuracy, privacy, and bias. Furthermore, Ferrara (2023) emphasizes the importance of encouraging clear and open discussions about AI in education, fostering transparency and accountability among stakeholders. The review of training data is essential to ensure the integrity

and reliability of AI-generated output, thereby mitigating concerns about factual inaccuracies and biases. In contrast, a more cautious approach entails seeking to limit or block access to GenAI tools altogether or simply refraining from their use in educational contexts (Selwyn, 2022). This approach acknowledges the inherent risks associated with AI tools and prioritises caution to prevent potential negative consequences. While intervention-based strategies focus on actively addressing concerns through policy interventions and transparency measures, the cautious approach adopts a preventive stance by avoiding potential risks associated with AI tools.

## 4. Research question and hypothesis

Our overarching research question is: *How does exposure to information about the prevalence of AI hallucinations and peers' ability to detect these hallucinations influence students' attitudes toward AI?*. We ground this question in management and information systems theory. The Technology Acceptance Model (TAM) posits that attitudes and intentions toward new technologies derive from perceived usefulness and ease of use David (1989)'s. Salient information about hallucinations can plausibly *lower* perceived usefulness ("if AI produces errors, its utility is reduced") and *raise* anticipated effort ("I must verify outputs"), together dampening attitudes. Research on trust in technology further suggests that error awareness typically depresses *confidence* and *reliance*, while sometimes fostering more calibrated, reflective forms of trust McKight (2002). In addition, drawing on capability theory, individuals' foundational academic and writing skills may shape how they translate risk signals into attitudes and behaviors (Sen, 1999; Nussbaum, 2011). We use this capability perspective to motivate our design and interpretation—particularly equity-sensitive pedagogy—without introducing a separate moderation hypothesis.

**Hypotheses.** Guided by these perspectives and our randomized exposure design, we test:

**H1 (Attitude dampening).** Random exposure to information about the prevalence of AI hallucinations and peers' low detection rates *negatively* affects students' attitudes toward AI, operationalized as **lower confidence in detecting AI errors** and **higher worry about hallucinations**.

**H2 (Learning demand).** Despite H1, exposure *increases* students' **preference for critical training on AI** (i.e., demand for instruction on risks and verification).

Constructs are mapped to measures as follows: confidence in AI error detection (trust calibration), worry about hallucinations (perceived risk/negative attitude), and preference for critical AI training (learning demand/competence-building). Our randomized design isolates the causal effect of exposure from confounds, aligning the empirical test with these theorized mechanisms. Equity considerations from capability theory inform our interpretation and motivate exploratory heterogeneity checks by prior academic performance to contextualize effect sizes.

## 5. Methodology

This research draws on data from a Year 2 econometrics course in a UK business research-led school. The course is compulsory for BSc Economics and BSc Economics and Management students, with a focus on linear regression, inference, and policy evaluation methods. Data were collected from coursework submissions and post-course surveys to examine how students' exposure to the prevalence of hallucinations of AI and peers' detection capabilities influence their attitudes towards AI. For further details of the questions and the process, see our companied paper in which we found 20% of the students fail to detect AI Hallucinations (Dang and Nguyen, 2025, mimeo).

The coursework, completed during Week 5, included a question requiring students to identify AI hallucinations in responses generated by ChatGPT3.5. Detailed marks for sub-questions from 211 students were recorded. After publishing solutions and feedback emphasizing AI hallucination risks, a

post-course survey was conducted two months later. The students were randomly divided into two groups: the "exposed" group received a factual statement about the low success rate of their cohort (20%) in identifying hallucinations due to AI, while the "non-exposed" group did not. After solutions and feedback highlighting AI hallucination risks were released, we implemented a *randomized information exposure* within the post-course survey. Students were randomly assigned to one of two versions: an *exposed* version that began with a factual statement noting that only a minority of peers successfully detected hallucinations in the coursework, and a *non-exposed* version without this statement. We measured three attitudinal outcomes on 1–4 Likert scales: worry about hallucinations, confidence in detecting AI errors, and preference for further critical training on AI. Participation was voluntary, incentivized with a 20 voucher, and achieved low response rates (23 from the exposed group and 20 from the non-exposed group). Table A1 in the Appendix describes the wording of the questions.

Primary outcomes are: (i) *Confidence in AI error detection* (higher = more confident), (ii) *Worry about hallucinations* (higher = more worried), and (iii) *Preference for critical AI training*. These map to theorised constructs of trust calibration, perceived risk, and learning demand. Where available, background indicators (e.g., prior coursework performance) are used descriptively and as precision covariates in robustness checks.

Accordingly, we interpret between-arm differences as *associational patterns consistent with* the theorised mechanism of information-induced trust calibration, rather than as definitive causal effects. Findings should be viewed as *suggestive evidence* motivating larger, pre-registered studies with higher response rates and tracked baselines. The randomised design aimed to isolate causal effects, though the small sample size limited statistical power and generalizability.

## 6. Findings

We find notable differences in students' attitudes toward AI based on their exposure to information about AI hallucinations. Among the 43 survey respondents, students who were randomly informed about the low rate (20%) of their peers successfully detecting AI hallucinations exhibited less confidence and greater caution regarding their understanding and ability to identify AI errors. In contrast, students who were not exposed to this information displayed higher confidence and were less concerned about the risks of hallucinations due to AI.

Interestingly, both groups expressed a strong interest in receiving additional training on the use and dangers of AI within their econometrics and economics curriculum, although a small minority of the exposed group disagreed with this sentiment. These results suggest that awareness of AI limitations and inaccuracies can lead to a more critical, albeit negative, perspective on AI, underscoring the importance of building foundational academic skills and fostering AI literacy in students' educational experiences.

Table 1 reports t-tests to compare the means of the four different variables of the not-exposed students (16) and randomly exposed students (20). Students answer whether they agree with the statement in a Likert scale from 1-4, with 4 being strongly agree and 1 being strongly disagree.

**Table 1**
Random exposure to information about peer's GenAI hallucination detection and Attitude towards GenAI

| Variables | Not Exposed | Exposed | Difference | Std. Error |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Understanding of Hallucinations | 2.00 | 3.25 | -1.25*** | 0.256 |
| Worrying about Hallucinations | 3.688 | 2.25 | 1.44*** | 0.341 |
| Confident about Detecting | 1.688 | 3.35 | -1.66*** | 0.257 |
| Prefer Critical Training on AI | 3.31 | 3.35 | -0.038 | 0.200 |
| Number of observations | 23 | 20 | | |

*Notes:* *p $\leq$ .10; **p $\leq$ .05; ***p $\leq$ .01. The table reports unconditional t-tests to compare the means of four different variables of the not-exposed students (16) and randomly exposed students (20). Students answer whether they agree with the statement on a Likert scale from 1–4, with 4 being strongly agree and 1 being strongly disagree.

Table 1 seems to paint a rather pessimistic view. Students who are randomly made aware of the rather poor performance in detecting AI hallucinations in the coursework seem to be less confident and more cautious about their understanding of AI hallucinations. In contrast, the randomly unaware students are less likely to worry about AI hallucinations, and more confident in their ability to detect AI errors. Both groups are almost equally interested in more training in AI hallucinations as part of their econometrics and economics curriculum, with a few aware students disagree with that statement. In essence, we observe a negative shift in students' confidence in their own understanding of AI hallucinations and their ability to detect the issues when being made aware of the prevalence of the issues and their peers' situations. Both groups all prefer to learn and get training more about the use and danger of AI in their degree. This result highlights the findings we found in previous sections: what causes students to be cautious and critical (negative) about AI is the critical knowledge of the probability that AI can make serious incorrectness and inconsistency. Together with the emphasis on academic knowledge, foundational subject materials, and writing skills, it is the awareness of AI and its potential problems that have become crucial components in the toolkits for students.

The findings highlight that awareness of AI hallucinations, and their prevalence can shift students' attitudes toward being more cautious and critical. This suggests that equipping students with knowledge about AI's capabilities and limitations is essential for preparing them to use such tools effectively in academic and professional contexts.

## 7. Implications

This research underscores the importance of explicitly embedding AI literacy into curricula, not only by raising awareness of AI hallucinations but also by equipping students with practical strategies to critically engage with AI-generated content. Educators should design authentic assessments that mirror real-world scenarios where professionals must evaluate information of varying reliability, thereby cultivating both subject knowledge and epistemic vigilance. For example, assignments could require students to triangulate AI outputs against academic literature, policy reports, or datasets, encouraging habits of verification and cross-checking.

At an institutional level, solution strategies could include:

- **Curricular integration**: Embedding modules or workshops on AI hallucinations and misinformation detection within core courses, rather than treating them as optional extras.
- **Faculty development**: Training educators to design assessments that leverage AI tools responsibly, shifting from a "ban or embrace" dichotomy towards structured critical engagement.
- **Equity considerations**: Providing targeted support for students with weaker academic writing or analytical foundations, ensuring that AI literacy does not exacerbate existing achievement gaps.
- **Cross-disciplinary collaboration**: Partnering with computer science and information studies to co-develop resources that address technical as well as ethical dimensions of AI hallucinations.

For further research, larger-scale studies are needed to test the generalizability of these findings across disciplines, institutions, and cultural contexts. Longitudinal research could also examine whether training interventions sustainably improve students' detection skills and confidence over time. Finally, comparative studies could explore whether awareness of AI errors leads to more critical but constructive adoption of GenAI tools in professional settings, such as consulting, journalism, or policy analysis.

## 8. Conclusions

This study demonstrates how students' awareness of AI hallucinations—and their peers' difficulties in detecting them—shapes attitudes toward AI. While exposure can reduce confidence and heighten caution, both exposed and non-exposed groups express strong demand for training on AI's risks and

capabilities. These findings suggest that critical engagement, rather than avoidance, should guide pedagogical approaches to GenAI in higher education.

Our results highlight three key takeaways. First, authentic assessment designs that challenge students to evaluate AI outputs can foster both subject mastery and digital literacy. Second, equity must remain central: without careful scaffolding, students with stronger academic foundations will benefit disproportionately from AI-enabled learning. Third, fostering critical AI literacy requires a balance—raising awareness of risks while equipping students with strategies for constructive, confident use.

By situating AI hallucinations not as a deterrent but as a learning opportunity, educators can prepare students to navigate the complexities of an AI-driven professional world with both caution and competence.

## Acknowledgments

We acknowledge generous funding from the Innovative Education Fund from King's Business School, and various participants in workshops and conferences. All errors are of our own. For the template, we thank the the developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart and to the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates.

## Declaration on Generative AI

The author(s) used ChatGPT-3.5 for creating the AI hallucination. We do use ChatGPT-4 for spelling check and suggestions to improve the writing. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

[1] Angelucci, C., & Prat, A. (2024). Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News. *American Economic Review*, 114(4), 887–925. https://doi.org/10.1257/aer.20211003.

[2] Amoozadeh, M., Daniels, D., Nam, D., Kumar, A., Chen, S., Hilton, M., Srinivasa Ragavan, S., & Alipour, M. A. (2024). Trust in Generative AI among Students: An Exploratory Study. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education (SIGCSE '24)* (pp. 1261–1267). https://doi.org/10.1145/3626252.3630842.

[3] Dang, C. T.,& Nguyen, A. (2025). Distinguishing Fact from Fiction: Student Traits, Attitudes, and AI Hallucination Detection in Business School Assessment. *arXiv:2506.00050*. https://arxiv.org/abs/2506.00050.

[4] Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *arXiv:2304.03738*. https://arxiv.org/abs/2304.03738.

[5] Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: global reactions to AI innovations. *Scientific Reports*, 13, 42227. https://doi.org/10.1038/s41598-023-42227-6.

[6] Holstein, K., McLaren, B. M., & Aleven, V. (2018). Mitigating Knowledge Decay from Instruction with Voluntary Use of an Adaptive Learning System *Proceedings of the International Conference on Artificial Intelligence in Education*, 119–133. https://doi.org/10.1007/978-3-319-93846-2_23.

[7] Hooda, D. S., Mittal, N., Gupta, S., & Saraswat, R. (2022). Artificial Intelligence for Assessment and Feedback to Enhance Student Success. *Mathematical Problems in Engineering*, 2022, 9933578. https://doi.org/10.1155/2022/5215722.

[8] Hyde, S., Busby, A., & Bonner, R. L (2024). Tools or Fools: Are We Educating Managers or Creating Tool-Dependent Robots? *Journal of Management Education*, 48(4), 708-734. https://doi.org/10.1177/10525629241230357.

[9] Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J. and Käser, T. (2025). The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions. *Computers and Education: Artificial Intelligence*, 8, p.100368. https://doi.org/10.1016/j.caeai.2025.100368.

[10] Selwyn, N. (2022). The Future of AI and Education: Some Cautionary Notes. *European Journal of Education*, 57(4), 620-631. https://doi.org/10.1111/ejed.12532.

[11] Stein, J. P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M.(2024). Attitudes towards AI: measurement and associations with personality. *Scientific Reports*, 14:1, 14(1), 1–16. https://doi.org/10.1038/s41598-024-53335-2.

[12] The Times. (2023, January 12). Universities move to curb ChatGPT use amid plagiarism fears. Retrieved from https://www.thetimes.co.uk/.

[13] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., & Chen, Y. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models *arXiv:2309.01219*. https://arxiv.org/abs/2309.01219.

00