Making thinking visible: AI-generated process mapping to enhance feedback processes in simulation-based learning

Suzanne Estaphan, School of Medicine and Psychology, Australian National University, Australia; Marcos Rojas, Graduate School of Education, Stanford University, USA; Gerry Corrigan, School of Rural Medicine, Charles Stuart University, Australia; Amanda K Edgar, Deakin University, Australia

Background

Clinical reasoning (CR) is a cornerstone capability for health professionals, yet its underlying cognitive processes are inherently difficult to observe and assess (Lee et al, 2013). Health professional education (HPE) programmes employ a range of pedagogical strategies, such as problem-based learning and clinical placements, to develop this capability (Edgar et al, 2023; Rudland et al, 2025). While these approaches support learners' engagement with clinical problems, they primarily emphasise observable performance outcomes rather than the cognitive processes that drive decision-making (Daniel et al, 2019).

Simulation-based education (SBE) is widely used to develop CR by placing students in realistic clinical scenarios that require decision-making, consequence evaluation and reflection (Durning and Artino, 2011; Edgar, Chong et al, 2024). However, in such situations, the process of CR is not visible for teachers, assessors and the students themselves (Audétat et al, 2017).

One approach to making these processes visible is process mapping. This method captures and visually represents CR in the form of maps (Corrigan, 2001) supporting more focused reflection and targeted feedback by making CR visible in SBE. Feedback in SBE is a core element of these learning activities and is often embedded as structured debriefing, where students explore the relationships among events, feelings and performances (Eddy et al, 2013; Fanning and Gaba, 2007; Rudolph et al, 2007). Whether facilitated or self-guided, this is a reflective element of SBE that is critical for learning (Issenberg et al, 2005; Motola et al, 2013). Process mapping could enhance this process; however, its implementation requires substantial time, resources and expertise, limiting its scalability in routine educational practice.

With the rapid emergence of artificial intelligence (AI), attention has shifted to its potential to support timely, scalable and personalised feedback (Chan and Hu, 2023; Lee and Moore, 2024; Shi and Aryadoust, 2024). Despite this growing interest, little is known about the feasibility and trustworthiness of AI-generated process mapping. Integrating AI-generated process mapping in SBE may yield novel opportunities to surface learners' clinical reasoning processes and enhance feedback processes while reducing resource burden.

This study aimed to explore the feasibility of developing AI-generated process mapping as a methodological approach. Specifically, it investigated the trustworthiness of AI-generated

annotations and visualisations of CR, using expert-generated outputs as a benchmark. It was guided by two research questions:

+ Can AI-generated annotations reliably identify components of CR in process mapping transcripts?
+ Can AI-generated process maps validly represent CR pathways in a way that aligns with expert judgement?

Approach

This case study drew on an established process mapping methodology that involved analysing interview transcripts and generating visual process maps. A combination of quantitative and qualitative approaches was used to assess the trustworthiness of AI-generated process maps. Reliability of outputs was assessed using statistical analysis, and validity was examined qualitatively through expert review.

The study was conducted in the context of a CR simulation, which has previously been used as a case example in research on process mapping methodology (Edgar, Estaphan et al, 2024). To focus on methodological trustworthiness, synthetic transcripts were used. This allowed full control over transcript structure and content, which is helpful when evaluating AI-generated annotation and process mapping. This approach is consistent with both ethical and common practice in AI development (Hao et al, 2024; Whitehouse et al, 2023).

To generate synthetic transcripts, a researcher simulated a student and responded to process mapping interview questions developed in a previous study. This initial transcript was then provided to GPT-4o1 (OpenAI, California, USA; hereafter referred to as the LLM) to generate synthetic transcripts. Three transcripts were generated, which aligns with prior methodological work using the same simulation assessment and number of transcripts. The research team then reviewed the transcripts and adjusted wording where needed to better reflect the clinical context – for example, changing *'measurement of refractive error'* to *'refraction'* (Kumar, 2024).

Next, the transcripts were annotated by both humans and the LLM using a structured codebook from a previous study (Edgar, Estaphan et al, 2024). Annotation work is typically an iterative coding process (Chun Tie et al, 2019; Glaser and Strauss, 2017). The LLM then used these annotations to generate Mermaid.js scripts, which were rendered in a compatible diagramming tool to produce visual process maps of the reasoning pathway. To support accurate AI-generated annotation and process mapping, the study followed published best practices, with the workflow summarised in Figures 1 and 2 (Birks et al, 2008; Edgar, Estaphan et al, 2024; Jiang et al, 2025; Törnberg, 2024).

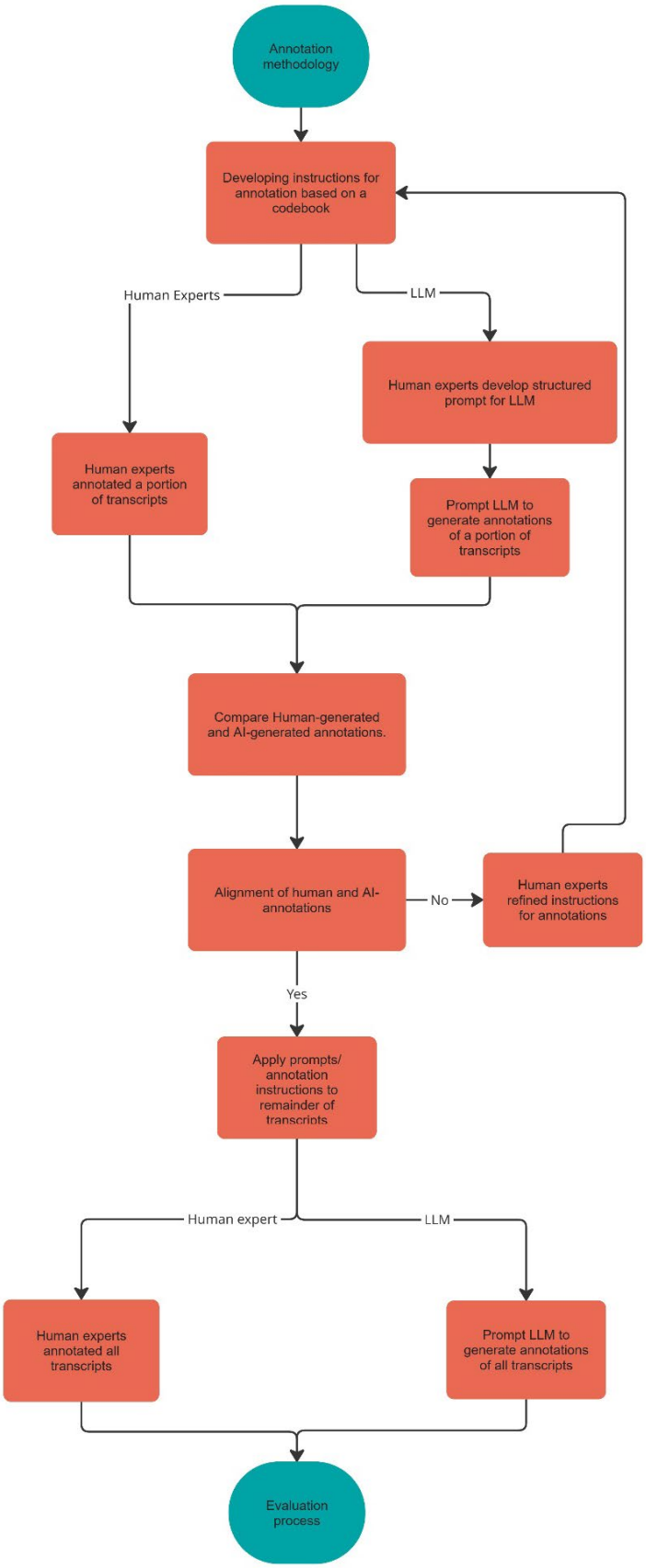

Figure 1. The steps for human and AI-generated annotations

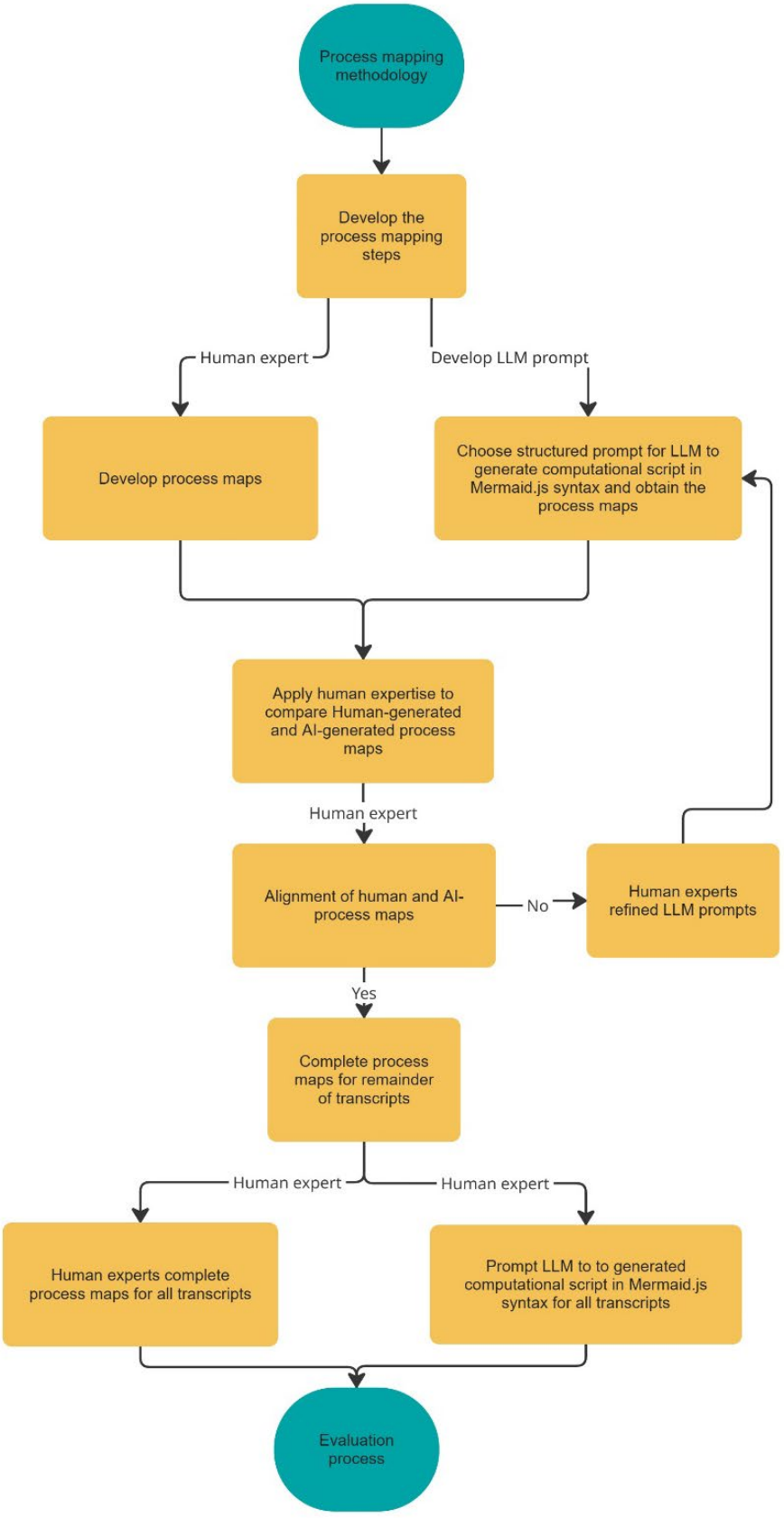

Figure 2. The steps for human and AI-generated process maps

To evaluate the reliability of AI-generated annotations, two types of comparison were conducted. R Studio was used to analyse agreement in the categorical data (RStudio Team, 2024). Intra-rater reliability was examined to determine whether the LLM produced consistent annotations when the same transcript was processed three times under identical conditions. This was assessed using percentage agreement (Ahmed and Ishtiaq, 2021; Munro, 2005). Inter-rater-reliability compared AI-generated annotations with human-generated annotations using Cohen's Kappa (κ) for alignment between AI and human-coded transcripts (Brennan and Prediger, 1981; Edgar et al, 2022).

Kappa values were interpreted as follows: near perfect (>0.81), substantial (0.61-0.80), moderate (0.41-0.60), fair (0.21-0.40), slight (0.00-0.20), and poor (<0.00) (Ranganathan et al, 2017). Discrepancies between annotations were analysed and discussed. For each transcript, Cohen's Kappa (κ) was calculated for every AI run, resulting in three κ values per transcript and a mean Kappa.

Because process mapping is still a new method in health professions education, the validity assessment was informed by related methodological guidance (Ahmed and Ishtiaq, 2021; Edgar, Estaphan et al., 2024; Markham et al, 1994). Process mapping and concept mapping both visually represent cognitive structures, although they are not identical methods. Given these similarities, the authors drew on approaches used in concept mapping research to assess validity in process mapping (Markham et al, 1994). Internal representational validity was assessed by a researcher with expertise in process mapping, who compared all AI-generated maps with the corresponding human-generated maps (Rosas and Kane, 2012). The aim was not to identify a single correct map, but to examine how closely the AI-generated maps matched the overall structure of the human-generated maps and whether they adequately represented participants' judgments.

Outcomes

Main findings

The main finding was that the AI could generate annotations and process maps that were often similar to those produced by human experts, although performance varied across transcripts. Across repeated runs, AI agreement was stronger for Transcripts 1 and 3 and notably weaker for Transcript 2, the shortest transcript (Table 2). A similar pattern was seen when AI annotations were compared with expert annotations: overall agreement was moderate, with stronger agreement for Transcripts 1 and 3 and lower agreement for Transcript 2 (Table 3).

Transcript	Percentage of annotation agreement (SD)	Percentage of sub-annotation agreement (SD)
Transcript 1	86.90 (29.17)	78.57 (35.39)
Transcript 2	37.50 (20.41)	23.61 (28.62)
Transcript 3	90.48 (23.61)	72.22 (34.47)
Mean	71.63 (32.79)	58.13 (27.48)

Table 2. AI agreement across the three runs and transcripts

Transcript	Annotation Kappa Mean (SD)	Subannotation Kappa Mean (SD)
Transcript 1	0.65 (0.05)	0.67 (0.05)
Transcript 2	0.29 (0.14)	0.31 (0.15)
Transcript 3	0.79 (0.03)	0.56 (0.07)
Overall	0.58 (0.25)	0.51 (0.18)
Note: Cohen's Kappa is computed across three runs per transcript.		

Table 3. Mean Cohen's Kappa between three AI runs and expert annotations by transcript

This variation is important. It suggests that AI-generated annotation may work better when transcripts are more developed and clearly structured, and less well when responses are shorter or more uncertain. The lower performance for Transcript 2 therefore does not weaken the study; rather, it provides a more realistic picture of where this approach may be more or less reliable.

Figure 3a shows the human-generated process maps for Transcript 1, and Figure 3b shows the corresponding AI-generated maps. Although some differences were present, the overall structure of the maps was similar, and both represented a reasoning pathway leading to a clinical decision. This supports the potential of AI-generated process maps as a way to make learners' reasoning more visible.

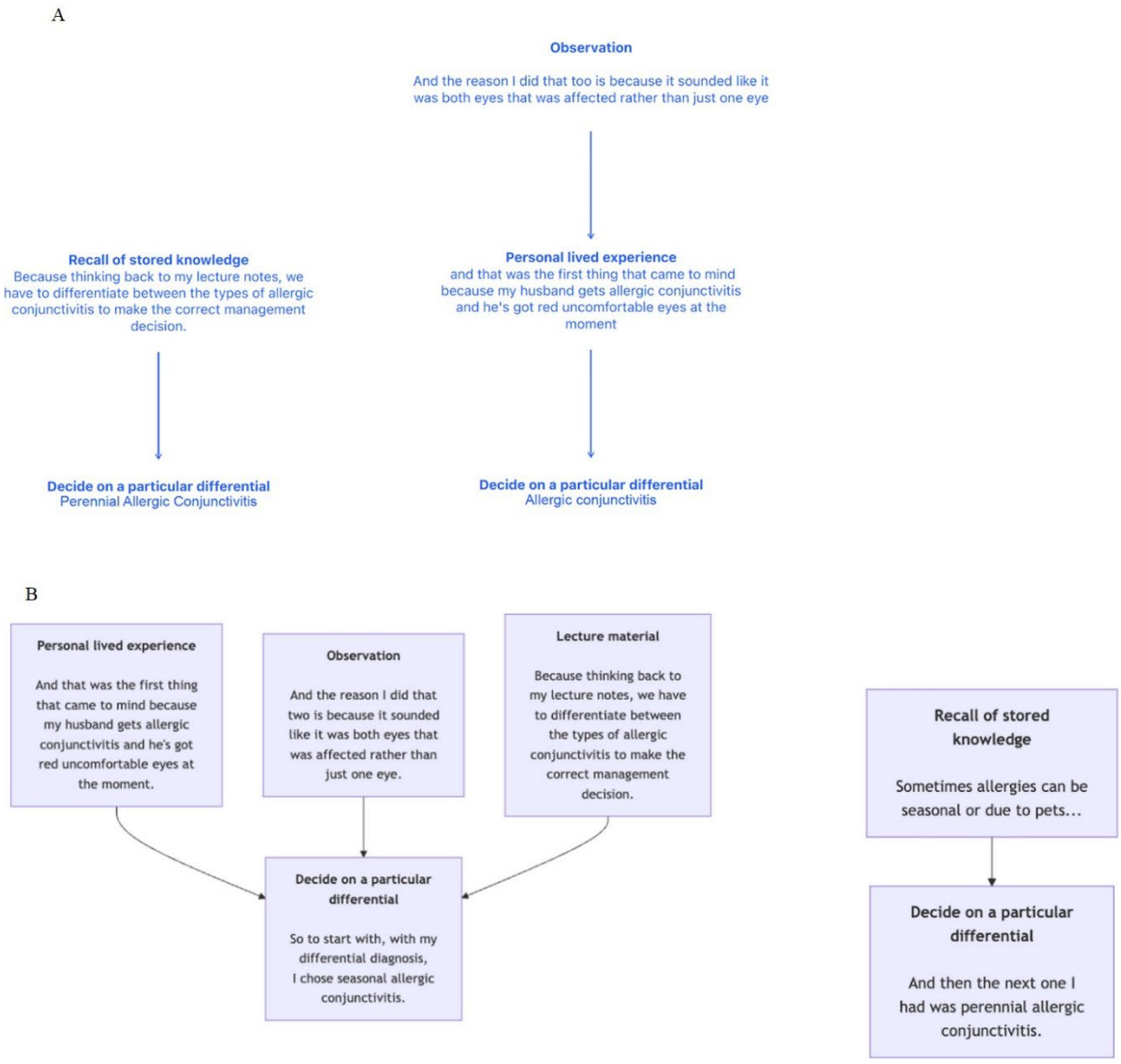

Figure 3a. Human-generated process mapping
Figure 3b. AI-generated process maps

Key messages and transferability

In conclusion, these results indicated moderate overall inter-rater reliability. Substantial agreement was achieved for Transcripts 1 and 3, whereas Transcript 2 showed only fair agreement. Of the three generated transcripts, the second, which was produced using a prompt specifying that the student had some level of uncertainty, was approximately half the length of the first and third transcripts. The short transcript had lower agreement in intra- and inter-rater analysis. This suggests that certain transcripts may inherently align better with AI processing, possibly due to the clarity and structure of the content or the cognitive process depicted in the transcript.

Future research could investigate optimal transcript length or structure, as well as explore whether variation in CR expertise among students influences AI performance. Regardless, as a first step, our results demonstrated that AI-generated annotations and visual mapping

were broadly comparable to those produced by experts confirming feasibility for further investigation of AI-generated process mapping integration into the SBE debriefing/feedback phase.

Institutions wishing to replicate this approach should also consider the time and expertise required. Although AI may reduce some manual work, implementation still depends on transcript preparation, prompt refinement and expert review of outputs. At this stage, the approach is best seen as supporting educator feedback rather than replacing it.

In practice, AI-generated maps could be integrated into the debriefing/feedback phase of SBE, to facilitate feedback that is detailed, timely and meaningful. Previous studies integrating AI with SBE have largely focused on assessing students' CR against a predefined framework, often classifying transmitted information (marks, comments, prompts) as feedback (Brügge et al, 2024; Furlan et al, 2022; Furlan et al. 2021). In contrast, process mapping makes explicit the student's own reasoning process, potentially enabling effective self-reflection, personalised learning and hence a deeper engagement with feedback. The aim is not to replace human judgement but to augment it by enhancing the visibility of the notoriously complex cognitive processes that underpin CR for both learners and educators. (Bearman and Luckin, 2020)

With the considered alignment to organisational policies and the ethical use of AI, exploring these threads further may uncover how to support learning CR with SBE. While relying solely on AI with the entirety of the feedback/debriefing phase presents ethical, pedagogical and technical challenges (Corbin et al, 2025), the value of AI may lie in "augmentation", rather than substitution, helping students to improve their capacity to learn.

References

Ahmed, I and Ishtiaq, S (2021) 'Reliability and validity: importance in medical research', *Methods*, 12 (1): 2401-2406. Available at: doi.org/10.47391/jpma.06-861

Audétat, M C, Laurin, S, Dory, V, Charlin, B and Nendaz, M R (2017) 'Diagnosis and management of clinical reasoning difficulties: part I Clinical reasoning supervision and educational diagnosis', *Medical Teachwe*, 39 (8): 792-796. Available at: doi.org/10.1080/0142159X.2017.1331033

Bearman, M and Luckin, R (2020) 'Preparing university assessment for a world with AI: tasks for human intelligence', in *Re-imagining university assessment in a digital world.* Cham: Springer, pp 49-63. Available at: doi.org/10.1007/978-3-030-41956-1_5

Birks, M, Chapman, Y and Francis, K (2008) 'Memoing in qualitative research: probing data and processes', *Journal of Research in Nursing*, 13 (1): 68-75. Available at: doi.org/10.1177/1744987107081254

Brennan, R L and Prediger, D J (1981) 'Coefficient kappa: some uses, misuses, and alternatives', *Educational and Psychological measurement*, 41 (3): 687-699. Available at: doi.org/10.1177/001316448104100307

were broadly comparable to those produced by experts confirming feasibility for further investigation of AI-generated process mapping integration into the SBE debriefing/feedback phase.

Institutions wishing to replicate this approach should also consider the time and expertise required. Although AI may reduce some manual work, implementation still depends on transcript preparation, prompt refinement and expert review of outputs. At this stage, the approach is best seen as supporting educator feedback rather than replacing it.

In practice, AI-generated maps could be integrated into the debriefing/feedback phase of SBE, to facilitate feedback that is detailed, timely and meaningful. Previous studies integrating AI with SBE have largely focused on assessing students' CR against a predefined framework, often classifying transmitted information (marks, comments, prompts) as feedback (Brügge et al, 2024; Furlan et al, 2022; Furlan et al. 2021). In contrast, process mapping makes explicit the student's own reasoning process, potentially enabling effective self-reflection, personalised learning and hence a deeper engagement with feedback. The aim is not to replace human judgement but to augment it by enhancing the visibility of the notoriously complex cognitive processes that underpin CR for both learners and educators. (Bearman and Luckin, 2020)

With the considered alignment to organisational policies and the ethical use of AI, exploring these threads further may uncover how to support learning CR with SBE. While relying solely on AI with the entirety of the feedback/debriefing phase presents ethical, pedagogical and technical challenges (Corbin et al, 2025), the value of AI may lie in "augmentation", rather than substitution, helping students to improve their capacity to learn.

References

Ahmed, I and Ishtiaq, S (2021) 'Reliability and validity: importance in medical research', *Methods*, 12 (1): 2401-2406. Available at: doi.org/10.47391/jpma.06-861

Audétat, M C, Laurin, S, Dory, V, Charlin, B and Nendaz, M R (2017) 'Diagnosis and management of clinical reasoning difficulties: part I Clinical reasoning supervision and educational diagnosis', *Medical Teachwe*, 39 (8): 792-796. Available at: doi.org/10.1080/0142159X.2017.1331033

Bearman, M and Luckin, R (2020) 'Preparing university assessment for a world with AI: tasks for human intelligence', in *Re-imagining university assessment in a digital world.* Cham: Springer, pp 49-63. Available at: doi.org/10.1007/978-3-030-41956-1_5

Birks, M, Chapman, Y and Francis, K (2008) 'Memoing in qualitative research: probing data and processes', *Journal of Research in Nursing*, 13 (1): 68-75. Available at: doi.org/10.1177/1744987107081254

Brennan, R L and Prediger, D J (1981) 'Coefficient kappa: some uses, misuses, and alternatives', *Educational and Psychological measurement*, 41 (3): 687-699. Available at: doi.org/10.1177/001316448104100307

Brügge, E, Ricchizzi, S, Arenbeck, M, Keller, M N, Schur, L, Stummer, W, Holling, M, Lu, M H and Darici, D (2024) 'Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial', *BMC Medical Education*, 24 (1): 1391. Available at: doi.org/10.1186/s12909-024-06399-7

Chan, C K Y and Hu, W (2023) 'Students' voices on generative AI: perceptions, benefits, and challenges in higher education', *International Journal of Educational Technology in Higher Education*, 20 (1): 43. Available at: doi.org/10.1186/s41239-023-00411-8

Chun Tie, Y, Birks, M and Francis, K (2019) 'Grounded theory research: a design framework for novice researchers', *SAGE open medicine*, 7: 2050312118822927. Available at: doi.org/10.1177/205031211882292

Corbin, T, Tai, J and Flenady, G (2025) 'Understanding the place and value of GenAI feedback: a recognition-based framework', *Assessment and Evaluation in Higher Education*, 50 (5): 718-731. Available at: doi.org/10.1080/02602938.2025.2459641

Corrigan, G J (2001) *Conceptual development, investigative skills and decisions about processes*. Doctoral dissertation, Faculty of Education, University of Sydney.

Daniel, M, Rencic, J, Durning et al (2019) 'Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance', *Academic Medicine*, 94 (6): 902-912. Available at: doi.org/10.1097/ACM.0000000000002618

Durning, S J and Artino, A R (2011) 'Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide no. 52', *Medical Teacher*, 33 (3): 188-199. Available at: doi.org/10.3109/0142159X.2011.550965

Eddy, E R, Tannenbaum, S I and Mathieu, J E (2013) 'Helping teams to help themselves: Comparing two team-led debriefing methods', *Personnel Psychology*, 66 (4): 975-1008. Available at: doi.org/10.1111/peps.12041

Edgar, A, Ainge, L, Backhouse, S and Armitage, J (2022) 'A cohort study for the development and validation of a reflective inventory to quantify diagnostic reasoning skills in optometry practice', *BMC Medical Education*, 22 (1): 536. Available at: doi.org/10.1186/s12909-022-03493-6

Edgar, A, Chong, L X, Wood-Bradley, R, Armitage, J A, Narayanan, A and Macfarlane, S (2024) 'The role of extended reality in optometry education: a narrative review', *Clinical and Experimental Optometry*, 108 (3): 258-267. Available at: doi.org/10.1080/08164622.2024.2366366

Edgar, A, Estaphan, S, Chong, L, Armitage, J, Ainge, L and Corrigan, G (2024) 'Making reasoning visible through process mapping in digitally simulated clinical reasoning assessments', *Focus on Health Professional Education: A Multi-Professional Journal*, 25 (4): 38-46. Available at: doi.org/10.11157/fohpe.v25i4.796

Edgar, A K, Armitage, J A, Chong, L X, Arambewela-Colley, N and Narayanan, A (2023) 'Breaking boundaries and opening borders by clicking into an inclusive virtual simulated learning environment', *Education and Information Technologies*. Available at: doi.org/10.1007/s10639-023-12369-1

Fanning, R M and Gaba, D M (2007) 'The role of debriefing in simulation-based learning. *Simulation in Healthcare*, 2 (2): 115-125. Available at: doi.org/10.1097/SIH.0b013e3180315539

Furlan, R, Gatti, M, Mene, R, Shiffer, D, Marchiori, C, Giaj Levra, A, Saturnino, V, Brunetta, E and Dipaola, F (2022) 'Learning analytics applied to clinical diagnostic reasoning using a natural language processing–based virtual patient simulator: case study', *JMIR Medical Education*, 8 (1): e24372. Available at: doi.org/0.2196/24372

Furlan, R, Gatti, M, Menè, R, Shiffer, D, Marchiori, C, Giaj Levra, A, Saturnino, V, Brunetta, E and Dipaola, F (2021) 'A natural language processing–based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: simulator development and case study', *JMIR medical informatics*, 9 (4): e24073. Available at: doi.org/10.2196/24073

Glaser, B and Strauss, A (2017) *Discovery of grounded theory: Strategies for qualitative research*. Abingdon: Routledge. Available at: doi.org/10.4324/9780203793206

Hao, S, Han, W, Jiang, T, Li, Y, Wu, H, Zhong, C, Zhou, Z and Tang, H (2024) *Synthetic data in AI: Challenges, applications, and ethical implications*. arXiv. Availbale at: arxiv.org/abs/2401.01629.

Issenberg, S B, McGaghie, W C, Petrusa, E R, Lee Gordon, D and Scalese, R J (2005) 'Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review', *Medical Teacher*, 27 (1): 10-28. Available at: doi.org/10.1080/01421590500046924

Jiang, Y, Ko-Wong, L and Valdovinos Gutierrez, I (2025) 'The feasibility and comparability of using artificial intelligence for qualitative data analysis in equity-focused research', *Educational Researcher*, 54 (3). Available at: doi.org/10.3102/0013189X251314821

Kumar, A (2024) 'The role of synthetic data in advancing AI models: opportunities, challenges, and ethical considerations', *Journal of Artificial Intelligence General Science*, *5*(1): 443-459. Available at: doi.org/10.60087/jaigs.v5i1.256

Lee, A, Steketee, C, Rogers, G D and Moran, M (2013) 'Towards a theoretical framework for curriculum development in health professional education', *Focus on Health Professional Education: A Multi-Professional Journal*, 14 (3): 70-83.

Lee, S S and Moore, R L (2024) 'Harnessing generative AI (GenAI) for automated feedback in higher education: a systematic review', *Online Learning*, 28 (3): 82-106.

Markham, K M, Mintzes, J J and Jones, M G (1994) 'The concept map as a research and evaluation tool: further evidence of validity', *Journal of Research in Science Teaching*, 31 (1): 91-101. Available at: https://doi.org/10.1002/tea.3660310109

Motola, I, Devine, L A, Chung, H S, Sullivan, J E and Issenberg, S B (2013) 'Simulation in healthcare education: a best evidence practical guide. AMEE Guide No. 82', *Medical Teacher*, 35 (10): e1511-1530. Available at: doi.org/10.3109/0142159X.2013.818632

Munro, B H (2005) *Statistical methods for health care research* (5th edn ed) Lippincott Williams and Wilkins.

Ranganathan, P, Pramesh, C and Aggarwal, R (2017) 'Common pitfalls in statistical analysis: measures of agreement', *Perspectives in Clinical Research*, 8 (4): 187-191.

Rosas, S R and Kane, M (2012) 'Quality and rigor of the concept mapping methodology: A pooled study analysis', *Evaluation and Program Planning*, 35 (2): 236-245.

Rudland, J, Ash, J, Barclay, L, Bloomfield, J, Bogossian, F, Byrne, K, Butler, M, Chur-Hansen, A, Clark, S and Edgar, A (2025) 'Rethinking clinical placements: A response to changing healthcare demands', *Focus on Health Professional Education: A Multi-Professional Journal*, 26 (1): 60-77.

Rudolph, J W, Simon, R, Rivard, P, Dufresne, R L and Raemer, D B (2007) 'Debriefing with good judgment: combining rigorous feedback with genuine inquiry', *Anesthesiology clinics*, 25 (2): 361-376. Available at: doi.org/10.1016/j.anclin.2007.03.007

Shi, H and Aryadoust, V (2024) 'A systematic review of AI-based automated written feedback research', *ReCALL*, 36 (2): 187-209. Available at: doi.org/10.1017/S0958344023000265

Törnberg, P (2024) *Best practices for text annotation with large language models*. arXiv. Available at: arxiv.org/abs/2402.05129.

Whitehouse, C, Choudhury, M and Aji, A F (2023) *LLM-powered data augmentation for enhanced cross-lingual performance*. *arXiv.* Available at: arxiv.org/abs/2305.14288