ORIGINAL ARTICLE

Validation of AI-Generated Toxicology Vignettes in Singapore: A Cross-Sectional Expert Review

ADELINE NGO[1,*], LOKESH KRISHNAJI KOLHE[1], VAIKUNTHAN RAJARATNAM[2]

[1]Department of Emergency Medicine, Woodlands Health, National Healthcare Group (NHG) Health, Singapore
[2]Department of Orthopaedic Surgery, Khoo Teck Puat Hospital, National Healthcare Group (NHG) Health, Singapore

Abstract

***Background*:** Generative artificial intelligence (AI) holds promise for medical education, yet the realism and contextual relevance of AI-generated toxicology vignettes in Southeast Asia are not well established. This study evaluated the face and content validity of vignettes produced by ChatGPT-4.0, to assess their plausibility and relevance for use in Singaporean emergency medicine education, training, and clinical decision support.
***Methods*:** Ten vignettes were generated using ChatGPT-4.0 in March 2025 and independently evaluated by five Singapore-based clinical toxicologists from four public hospitals. A six-domain rubric, adapted from established validity frameworks, scored presentation realism, typicality of exposure, toxidrome representation, clinical progression, appropriateness for toxicology consultation, and alignment with local practice. Inter-rater reliability was calculated using a two-way random-effects intraclass correlation coefficient [ICC(2, k)].
***Results*:** The mean total score was 20.1/24 (SD = 1.8). Inter-rater agreement was excellent (ICC = 0.87; 95% CI: 0.80–0.94). Face validity averaged 4.4/5 (SD = 0.5) and content validity averaged 4.2/5 (SD = 0.6). Most vignettes reflected common regional poisoning patterns, with some depicting rare but plausible exposures relevant to local practice.
***Conclusion*:** ChatGPT-4.0 can generate toxicology vignettes with high expert-rated realism and contextual relevance when tailored to Singaporean practice. These findings support its potential role in medical education, simulation, and decision-support tools. Further research should compare AI-generated and clinician-authored materials to determine educational impact and applicability in real-world clinical settings.

Keywords: Toxicology; Artificial Intelligence; Clinical Simulation; Educational Measurement; Emergency Medicine

How to cite this article: Ngo A, Kolhe LK, Rajaratnam V. Validation of AI-Generated Toxicology Vignettes in Singapore: A Cross-Sectional Expert Review. Asia Pac J Med Toxicol 2025; 14(4): 140-4.

INTRODUCTION

Acute poisoning remains a significant challenge for emergency departments (EDs) worldwide, contributing to substantial patient morbidity, resource strain, and healthcare costs. In Singapore, EDs manage more than 3,000 toxic exposures annually across three public hospitals (2001–2003) [1], underscoring the need for timely recognition and management in high-throughput environments. Physicians have expressed a clear need for timely access to drug and poison information services, as limitations in availability can contribute to variability in consultation practices and delays in optimal management [2].

AI-enabled tools, including large language models, can help address gaps in clinical expertise by generating realistic educational and clinical scenarios across multiple domains [3]. In toxicology, such vignettes may support clinical decision-making, standardise training content, and serve as formative or summative assessment material. However, their educational and clinical utility depends on perceived authenticity, contextual relevance, and alignment with real-world practice, particularly in regions with unique exposure patterns.

Under Messick's unified framework of construct validity, authenticity relates to the degree to which an assessment reflects real-world complexity and context. In applied research, this often involves evaluating both content validity (accuracy of representation) and face validity (perceived

*Correspondence to: Dr. Adeline Ngo, MD. Department of Emergency Medicine, Woodlands Health, NHG Health, Singapore.
Email: angosy@gmail.com, Tel: +65 8228 8637

realism to expert users) [4, 5]. In toxicology, this requires precise toxidrome representation, inclusion of locally prevalent substances, and presentation of realistic ED decision prompts. While prior studies indicate that ChatGPT can generate high-quality toxicology content [6] and that AI tools can perform comparably to human toxicologists in simulated cases [7], these evaluations have been conducted in non-local contexts. No prior study has assessed the realism or regional relevance of AI-generated toxicology vignettes using input from practising clinical toxicologists in Singapore.

This study aimed to evaluate the face and content validity of ChatGPT-4.0–generated toxicology vignettes within the context of Singapore's ED practice.

METHODS

Theoretical Framework

This study followed Messick's unified theory of construct validity, focusing on:

- Face validity: Expert perceptions of realism and plausibility of clinical scenarios [4, 8].
- Content validity: The degree to which vignettes reflect essential toxicologic features [5, 8].

Study Design

A cross-sectional expert-rating study was conducted.

Rater sampling and blinding

Inclusion criteria were: (i) ≥ 2 years of clinical experience managing poisoning, overdose, or envenomation; and (ii) current employment at a Singapore healthcare institution.

Exclusion criteria were: (i) direct involvement in the development or commercialisation of AI technologies; and (ii) non-clinical roles without active patient care.

Invitations were distributed via institutional email. Raters performed their assessments independently and were blinded to one another's ratings and feedback throughout the study. Raters were blinded to any system-internal metadata (e.g., model parameters, retrieval sources, or prompt engineering details); only the AI-generated consult text was provided.

Participants

Five clinical toxicologists were recruited via purposive sampling from four public hospitals in Singapore to ensure domain expertise [9]. All participants had between 2 and over 10 years of toxicology experience and provided electronic informed consent. Reviewers were blinded to the fact that the cases were AI-generated until after all ratings were completed, and rated cases independently without discussion. No examples of pre-scored vignettes were provided to avoid anchoring bias.

Vignette Generation

Ten toxicology vignettes were generated in March 2025 using ChatGPT-4.0 (OpenAI, San Francisco, USA). The model was accessed via the ChatGPT Plus web interface (model version GPT-4.0, March 2025 update). Prompts were constructed to request realistic ED toxicology scenarios relevant to Singapore, including patient demographics, exposure history, examination findings, investigations, and key management steps. A complete example of the base prompt, with model parameters (temperature, token limits), is provided in Supplementary Appendix 1. No post-generation edits were made except to remove institution-specific names. Cases were designed to include both familiar and less frequent but plausible local exposures.

Evaluation Instrument

A six-domain rubric was adapted from prior educational validity tools [4, 5], with additional contextual considerations informed by recent vignette evaluation literature [10]. The domains are: (1) Presentation Realism; (2) Typical Toxin or Exposure; (3) Toxidrome & Clinical Course Accuracy; (4) Consult Prompt Relevance; (5) Appropriateness for Toxicologist; and (6) Fit with Singaporean Practice. Compared with the source rubric [4, 5], "Fit with Singaporean Practice" was added, and "Clarity of Learning Objective" was removed. The complete rubric, with rating definitions, is provided in Supplementary Appendix 2. A scoring sheet template is presented in Supplementary Appendix 3. The rubric was piloted with two toxicologists before data collection.

Data Collection and Analysis

Ratings were collected anonymously in March 2025. Descriptive statistics were reported. Inter-rater reliability was calculated using ICC(2,k) with Koo & Li thresholds [11]. Analyses were conducted in R version 4.3.3 using the psych package (v2.4.3). Standard error of measurement (SEM) assessed precision.

Ethics

No patient-identifiable data were used, and the study was conducted in accordance with the Declaration of Helsinki.

RESULTS

Ten AI-generated toxicology vignettes were evaluated, with all experts completing 100% of ratings.

Face and Content Validity

All vignettes received high ratings across domains (table 1). The mean total score was 20.1/24 (SD = 1.8). The highest-scoring domain was Appropriateness for Toxicologist (M = 3.88, SD = 0.43), while the lowest was Typical Toxin or Exposure (M = 2.69, SD = 0.72). Face validity (mean = 4.4/5, SD = 0.5) and content validity (mean = 4.2/5, SD = 0.6) were high. Some vignettes involved rare but plausible substances in Singapore. Expert comments indicated that realism was preserved despite infrequent local usage.

Inter-Rater Reliability

ICC(2,k) for total vignette scores was 0.87 (95% CI: 0.80–0.94), indicating excellent agreement among raters [11].

Comparative Case Scores

Highest scoring vignettes: Case 7 (22.2/24), Case 1 (21.6/24). Lowest scoring vignettes: Case 10 (17.6/24), Case 3 (19.4/24).

Table 1. Domain-Level Ratings for All Vignettes

Domain	Mean (SD*)	SEM**
Presentation realism	3.07 (0.78)	0.11
Typical toxin or exposure	2.69 (0.72)	0.10
Toxidrome & clinical course accuracy	3.07 (0.58)	0.08
Consult prompt relevance	3.76 (0.52)	0.07
Appropriateness for toxicologists	3.88 (0.43)	0.06
Fit with Singaporean practice	3.71 (0.56)	0.08

*SD: Standard deviation; **SEM: Standard error of measurement

DISCUSSION

This study evaluated the face and content validity of toxicology case vignettes generated by ChatGPT 4.0 (March 2025 release), focusing on applicability to Singaporean emergency medicine practice. Ten AI-generated cases were independently assessed by five clinical toxicologists from different public hospitals in Singapore, using a structured rubric encompassing presentation realism, typicality of exposure, toxidrome representation, clinical progression, appropriateness of toxicology consultation, and alignment with local workflows and therapeutic availability.

The vignettes received consistently high ratings (mean total score 20.1/24, SD = 1.8). Inter-rater reliability, measured using ICC(2,k), was 0.87 (95% CI: 0.80–0.94), signifying excellent consistency. The SEM was 0.29, reflecting high rating precision. Face validity (mean = 4.4/5) and content validity (mean = 4.2/5) were strong.

Unlike earlier studies conducted in non-local contexts [6, 7], this work emphasised clinical authenticity within a Southeast Asian ED setting, incorporating region-specific exposures and practice patterns. Reviewer comments noted that while some vignettes included substances or antidotes not routinely used in Singapore, they remained plausible, underscoring the importance of adapting LLM outputs to the specific clinical and organisational context of each healthcare system [12, 13].

These findings add to the literature on generative AI in medical education and training [14, 15]. In newly established or resource-constrained institutions, AI-generated cases could supplement training resources and standardise exposure to diverse scenarios. Future workflows may incorporate a Human-in-the-Loop (HITL) model, where AI-generated vignettes serve as a first draft, refined by expert educators to accelerate case development while preserving clinical accuracy.

With further validation, such content might be incorporated into decision-support tools, provided safeguards ensure accuracy, contextual fit, and clear delineation of AI involvement.

Importantly, these results represent an initial step rather than proof of clinical readiness. A planned second phase, approved by the National Healthcare Group (NHG) Ethics and Compliance Online System (ECOS), will directly compare AI-generated responses to these vignettes against those of human toxicologists to determine whether AI can complement expert judgment and to define appropriate use boundaries.

The small number of vignettes and reviewers limited this study. While inter-rater reliability was high, generalisability is constrained. The vignettes were designed for Singaporean emergency medicine practice; applicability to other Southeast Asian or non-Singapore contexts may be limited without adaptation to local poisoning patterns, antidote availability, and ED workflows. Expert perception was the primary outcome; objective measures of educational benefit or clinical decision-making impact were not assessed. Inclusion of rare but plausible exposures may have influenced "Typical Toxin or Exposure" scores. Although reviewers were blinded to AI authorship, the small specialist pool may introduce familiarity bias. This study utilized the ChatGPT-4.0 model (March 2025), representing a fixed version of the technology; subsequent updates may influence performance and reproducibility. Future studies should evaluate AI-generated content in larger and more diverse settings and compare performance directly against clinician-authored materials.

CONCLUSION

ChatGPT 4.0 can generate clinically plausible toxicology vignettes with high face and content validity when reviewed by Singaporean experts. These vignettes have potential for integration into education, training, and decision support where toxicology expertise is limited. Further research should compare AI-generated and clinician-authored management in real clinical settings.

Ethics Approval:

This study involved the expert review of AI-generated clinical vignettes for educational validation purposes. No patient-identifiable data were used, and no interventions involving human or animal subjects were performed. In accordance with local institutional policy for minimal-risk educational research, formal ethics committee review through the National Healthcare Group (NHG) Ethics and Compliance Online System (ECOS) was not required for this initial validation phase. All participants provided informed consent. The planned follow-up study (Phase 2), which will compare AI-generated and clinician-authored responses to these vignettes, has received ECOS approval (Reference: 2025-0347).

ACKNOWLEDGMENTS

We thank the clinical toxicologists who participated in this study for their time and expertise.

AI Use Disclosure: We acknowledge the use of ChatGPT 4.0 (OpenAI) for the generation of toxicology vignettes used in this research. The manuscript was authored by the study investigators.

Conflict of interest: None declared.

Funding and Support: No external funding was received.

REFERENCES

1. Ponampalam R, Tan HH, Ng KC, Lee WY, Tan SC. Demographics of toxic exposures presenting to three public hospital emergency departments in Singapore 2001–2003. Int J Emerg Med. 2009;2(1):25-31. doi:10.1007/s12245-008-0080-9
2. Ponampalam R, Anantharaman V. The need for drug and poison information—the Singapore physicians' perspective. Singapore Med J. 2003;44(5):231-42.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K. Large language models in medicine. Nat Med. 2023;29:1930–40. doi:10.1038/s41591-023-02448-8
4. Messick S. Validity of psychological assessment. Am Psychol. 1995;50(9):741-9. doi:10.1037/0003-066X.50.9.741
5. Rubio DM, Berg-Weger M, Tebb SS, Lee ES, Rauch S. Objectifying content validity: Conducting a content validity study in social work research. Soc Work Res. 2003;27(2):94–104. doi:10.1093/swr/27.2.94
6. Nogué-Xarau S, Ríos-Guillermo M, Amigó-Tadín M. Comparing answers of artificial intelligence systems and clinical toxicologists to questions about poisoning: Can their answers be distinguished? Emergencias. 2024;36(5):351–58. doi:10.55633/s3me/082.2024
7. Chary M, Boyer EW, Burns MM. Diagnosis of acute poisoning using explainable artificial intelligence. Comput Biol Med. 2021;134:104469. doi:10.1016/j.compbiomed.2021.104469
8. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e7-16. doi:10.1016/j.amjmed.2005.10.036
9. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health. 2015;42(5):533-44. doi:10.1007/s10488-013-0528-y
10. Bakkum MJ, Hartjes MG, Piët JD, Donker EM, Likić R, Sanz E, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Br J Clin Pharmacol. 2024;90(3):640-8. doi:10.1111/bcp.15977
11. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-63. doi:10.1016/j.jcm.2016.02.012
12. Wells C, Wollack J. An instructor's guide to understanding test reliability [Internet]. Madison: Testing & Evaluation Services, University of Wisconsin-Madison; 2003 [cited 2025 Aug 9]. Available from: https://testing.wisc.edu/Reliability.pdf
13. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI. Fine-tuning large language models for specialized use cases. Mayo Clin Proc Digit Health. 2025;3(1):100184. doi:10.1016/j.mcpdig.2024.11.005
14. Castano-Villegas N, Villa MC, Monsalve Barrientos K, Llano I, Zea J. Arkangel AI, OpenEvidence, ChatGPT, Medisearch: are they objectively up to medical standards? A real-life assessment of LLMs in healthcare [Preprint]. 2025. https://doi.org/10.1101/2025.09.23.25336206
15. Seo J, Choi D, Kim T, Cha WC, Kim M, Yoo H, et al. Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study (Preprint). J Med Internet Res. 2024. https://doi.org/10.2196/58329

Supplementary Appendix

Appendix 1 – Sample AI Prompt

Model: ChatGPT-4.0 (March 2025 version)

Access platform: ChatGPT Plus (OpenAI) web interface

Temperature: 0.7 | Max output tokens: 1,200 | Top p: 1.0 | Freq/Preset penalties: 0

Base prompt example:

"Generate a realistic emergency department toxicology case vignette suitable for a Singapore clinical setting. Include: patient age, gender, and relevant background; exposure history (substance name, route, timing, context); examination findings; laboratory results; clinical course over 4–8 hours in the ED; and key decision points prompting toxicology consultation. Avoid identifiable patient information or hospital names. Ensure the scenario is medically plausible within Singapore's healthcare context, with locally available investigations and treatments."

Note: No post-generation edits were made except for the removal of institution-specific names.

Appendix 2 – Six-Domain Rubric for Vignette Evaluation

Domain 1 – Presentation Realism

1 = Not realistic; implausible presentation … 4 = Highly realistic; fully consistent with clinical practice

Domain 2 – Typical Toxin or Exposure

1 = Exposure not seen in local practice … 4 = Commonly seen in local practice

Domain 3 – Toxidrome & Clinical Course Accuracy

1 = Inaccurate toxidrome/progression … 4 = Accurate and complete toxidrome/course

Domain 4 – Consult Prompt Relevance

1 = Unlikely to trigger consult … 4 = Strong and appropriate consult trigger

Domain 5 – Appropriateness for Toxicologist

1 = Not relevant to toxicology expertise … 4 = Highly relevant

Domain 6 – Fit with Singaporean Practice

1 = No alignment with local workflows/resources … 4 = Fully aligned

Appendix 3 – Template Scoring Sheet

Vignette ID	D1	D2	D3	D4	D5	D6	Total	Comments
Example	3	4	4	3	4	3	21	Notes…

Reviewer instructions: Rate each domain 1–4 per Appendix 2; provide brief comments; complete independently without discussion.