

Evaluating ChatGPT's performance on 100 forensic psychiatry and psychology multiple-choice questions (MCQs) reveals high surface-level accuracy but significant gaps in understanding complex psychological theories, legal procedures, and nuanced conceptual knowledge.

Evangelou K ¹, Pavlidou L ¹, Belimezakis N ¹, Hadjiphilippou M ², Louka C ³

¹ Department of Medicine, National and Kapodistrian University of Athens, Athens, Greece
² Department of Nursing, National and Kapodistrian University of Athens, Athens, Greece
³ Department of Chemistry, University of Cyprus, Nicosia, Cyprus

INTRODUCTION

Artificial intelligence (AI) and large language models (LLMs) are increasingly used in forensic psychiatry and psychology for education, decision support and documentation. However, their domain-specific accuracy and reasoning in structured assessment formats remain insufficiently evaluated. No study has yet examined the performance of the latest ChatGPT-5 model on targeted forensic psychology and forensic psychiatry multiple-choice questions (MCQs).

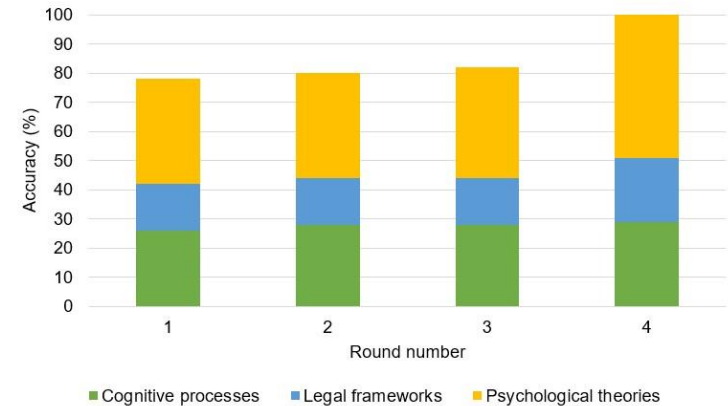
AIM

To evaluate the accuracy, error patterns and apparent “learning” across repeated rounds when ChatGPT-5 answers a validated set of forensic psychology and forensic psychiatry MCQs, and to explore implications for educational and examination use.

METHODS

We conducted a TRIPOD-LLM-adherent evaluation of off-the-shelf ChatGPT-5 (August 2025 version) on 100 MCQs from “Chapter 21: Forensic Psychology; 100 MCQs” (Hewstone, Wiley-Blackwell). Items covered foundations/history, eyewitness memory and testimony, interrogation and confession, investigative

interviewing, deception detection, offender profiling, criminology versus psychology, the Cambridge Study in Delinquent Development, developmental models of violence, and treatment approaches. Ninety-five (95) questions were single-best-answer and five were multi-select. The model received the verbatim English questions and options and was instructed to choose the best answer without additional prompt engineering. Four (4) consecutive rounds were run on the same item set; in each new round the model was shown its previous answer and the correct option was not disclosed. Responses were graded against the official answer key. Errors were classified into three domains (cognitive processing, legal frameworks, psychological theories) and descriptively analysed across rounds.



RESULTS

In round 1, ChatGPT-5 achieved 78.0% accuracy (78/100), with 22 errors: psychological theories 59.1%, legal frameworks 27.7%, cognitive processes 13.6%. Wrong answers clustered in theoretically dense or longitudinally complex content: Three foundational/historical questions (e.g. Hugo Münsterberg's contributions; forensic, criminological and legal psychology distinction), 6 on eyewitness memory and testimony (effects of leading questions, child versus adult witnesses, individual-difference predictor strength), 3 on criminology–psychology boundaries and juvenile crime, 4 questions arising from the Cambridge Study in Delinquent Development (predictive / non-predictive

risk factors, adult outcomes), and 6 on developmental and social-information-processing models of violence. Across rounds 2 and 3, the same domains remained problematic, with many errors simply “shifting” between distractors rather than converging on the correct concept, indicating unstable internal representations of theoretical constructs and risk factors. Although round-2 and round-3 accuracies increased modestly to 80.0% and 82.0%, respectively, several misunderstandings persisted (e.g. misclassification of distal antecedents, maintenance factors in violent behaviour, and key steps in social information processing) until the final remaining option was selected in round 4.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Cognitive processing

Legal frameworks

Psychological theories

The eventual 100% score therefore primarily reflects exhaustive elimination of previously tried options rather than demonstrable conceptual learning.

CONCLUSION

ChatGPT-5 showed high but imperfect single-round performance on specialised forensic psychiatry-psychology MCQs, with particular difficulty in psychological theories and nuanced research findings. Although accuracy improved across repeated rounds, the final perfect score reflected answer elimination rather than genuine reasoning gains. Forensic educators and examiners should therefore treat LLM outputs as adjunctive learning aids rather than standalone evaluators, avoid allowing multiple AI attempts on identical item pools, and maintain expert oversight. Larger, jurisdiction-diverse item sets, psychometrically characterised, are hence needed in order to define safe accuracy thresholds, and to guide responsible integration of LLMs into forensic mental health education and assessment.

CORRESPONDENCE

Kyriacos Evangelou; National and Kapodistrian University of Athens; NKUA
evangeloukyriacos@gmail.com