Advance online publication https://doi.org/10.5204/lthj.4896

Legal Minds Vs. Neural Networks: Law Schools Need to Confront Generative AI's Increasing Sophistication in Legal Reasoning

Armin Alimardani
Western Sydney University, Australia
Melissa Porter
University of Wollongong, Australia
Warwick Gullett
University of Wollongong, Australia

Abstract

This article reports on a follow-up investigation to a 2023 study examining the academic performance of Generative AI (GenAI) in university-level Criminal Law assessment. We evaluated GenAI models from five artificial intelligence (AI) providers across two compulsory subjects in the Bachelor of Laws (LLB) degree at the University of Wollongong: Criminal Law and Procedure A ('Criminal Law') and Law of Torts ('Tort Law'). For each subject, nine exam responses were generated by GenAI (n=18). Twelve of the 18 GenAI responses were blind marked by tutors who were not aware of AI involvement. The remainder were marked by the subject coordinators (authors of this study). The prompts used to generate the AI answers did not include subject-specific legal materials.

GenAI achieved mean scores of 76.3% in Criminal Law and 66.0% in Tort Law, outperforming on average 82.5% and 61.0% of students, respectively. Notably, seven of the 18 AI-generated papers ranked at or above the 90th percentile of students' performance. Compared with the 2023 baseline (mean 52.5%; percentile 22.1%), these results reflect substantial improvement in GenAI's academic performance within two years.

This study proposes a mixture of three assessment formats to ensure students demonstrate fundamental legal knowledge and skills and develop the capacity to collaborate with GenAI: (i) embedding GenAI as a legitimate component of assessment tasks; (ii) shifting written assignments, including mid-semester tasks, into AI-free invigilated formats; and (iii) a 'relay' approach where students either write a first draft and then use AI to analyse and suggest improvements or students prompt AI to write the first draft which students then review to verify claims and improve the output.

***Keywords**:* GenAI; ChatGPT; critical thinking; higher education; legal profession.

1. Introduction

A university degree is a claim about a person's capabilities. Conferral of a law degree does not certify that a graduate can produce fluent legal prose. Rather, it attests that the graduate can identify legal issues in interpersonal scenarios, select and weigh authority, reason by analogy, apply law to facts, verify their sources, and exercise judgement under professional constraint.[1] Those capacities are what an employer pays for and what a client relies upon. They are formed slowly through the very reading, drafting, and revision that generative artificial intelligence (GenAI) now performs in seconds.[2] This is why GenAI confronts legal education as a question of substance, not merely of academic integrity. If a model can generate work that passes

[1] Bell, "Becoming, Doing, Being"; Choi, "AI Assistance in Legal Analysis."
[2] Bell, "Becoming, Doing, Being."

for a competent graduate's, the degree's central promise is harder to keep, and it is more difficult to distinguish the graduate from the tool. Automation is reshaping entry-level legal work and rendering the graduate's position in the labour market less secure.[3] Deciding which capabilities a law degree must still guarantee, how to teach them, and how to assess them once a machine can imitate their outward form is therefore not a question to be settled by intuition or by vendors' marketing. It requires evidence of what these systems can and cannot actually do in the assessments law schools use to certify competence. This evidence is needed to form a defensible foundation upon which a revised legal education curriculum and policy can be built.[4] Emerging legal education scholarship has begun to move beyond academic integrity concerns to examine how GenAI tools might be incorporated into teaching, learning, and formative assessment.[5] When GenAI first entered this debate, the prevailing assessment was reassuring. Early legal-domain testing found the models distinctly limited. They fabricated authorities, reasoned from a fixed knowledge cut-off, and transferred poorly between jurisdictions. Systematic measurement has since confirmed how deep the reliability problem runs. For instance, studies from 2024 and 2025 showed that assessment of verifiable questions about a randomly selected United States (US) federal case led general-purpose models to hallucinate between 58% and 88% of the time and they were poorly calibrated when doing so.[6] Even purpose built legal research systems claimed as 'hallucination-free' returned unfounded answers in a sizeable minority of queries.[7]

In 2023, an empirical study by Armin Alimardani provided an Australian benchmark by generating 10 GenAI responses to a real Criminal Law and Procedure examination and having them blind marked alongside 225 student papers.[8] It was revealing on three counts. First, it showed that non-invigilated assessment was vulnerable. Some copy-and-paste exam questions without prompting resulted in a pass, and none of the tutors detected the artificial intelligence (AI) papers during marking. Second, headline claims about GPT-4's performance in the top 10% of US bar examination takers travelled poorly to an Australian university assessment.[9] Most AI papers fell in the bottom 30% of students, and the strongest reached roughly the 86th percentile only after substantial prompt engineering and subject-specific grounding. Third, it located a possible human advantage: GenAI performed comparatively well on simpler essay-style questions but struggled with the critical legal analysis demanded by hypothetical problem scenarios, especially the application stage of structured legal reasoning. The study identified an integrity risk and a pedagogical boundary, with applied legal reasoning as the area where students still held a meaningful edge.

This article reports a follow-up study designed to test whether that apparent boundary still holds. We evaluated models from five AI providers across two compulsory subjects, Criminal Law and Procedure A ('Criminal Law') and Law of Torts ('Tort Law'). Nine exam responses were generated for each subject (18 in total), with 12 blind marked by tutors who did not know about AI involvement and the remainder by subject coordinators who did know. Crucially, and unlike some of the 2023 prompts, the 2025 prompts provided no subject-specific teaching materials.

Across the two subjects, GenAI models achieved substantial improvement within two years compared with the 2023 Criminal Law baseline. The findings suggest that legal education can no longer rely on the assumption that students retain a comfortable edge over GenAI. Most importantly, the gap between AI and students on applied legal reasoning has narrowed. In Criminal Law, AI performance on problem and essay-style questions was almost identical, and in Tort Law, AI was near parity with students on problem questions and dramatically stronger on the essay. The critical analysis once identified as GenAI's defining weakness is now uneven rather than absent. This is not to say the models are reliable legal reasoners. Stronger reasoning appeared alongside marked inconsistency between and within models,[10] uneven citation practices, and hallucinated or unsupported authorities, confirming that verification remains a significant competency of legal AI literacy.[11]

The implication is not that assessment is futile but that it must be redesigned while a meaningful human advantage remains to be protected. We argue that law schools need an assessment architecture that secures foundational legal skills while teaching students to work with GenAI productively and critically. It is this combination upon which graduates' employability now depends. Part of that architecture should remain AI-free and invigilated. This is not a nostalgic retreat to pre-AI examining, but

[3] Madden, "The Dignity of Legal Education vs Artificial Intelligence in Australian Law Schools."
[4] Borges, "Could ChatGPT Get an Engineering Degree?"; Lodge, "Mapping out a Research Agenda for Generative Artificial Intelligence in Tertiary Education"; Thompson, "AI in Tertiary Education."
[5] See Hargreaves, "ChatGPT and Other Generative AI Tools as University Teaching Aids."
[6] Dahl, "Large Legal Fictions." See also Hargreaves, "'Words Are Flowing Out Like Endless Rain Into a Paper Cup'"; Burgess, "Using Generative AI to Identify Arguments in Judges' Reasons."
[7] Magesh, "Hallucination-Free?"
[8] Alimardani, "Generative Artificial Intelligence vs. Law Students."
[9] OpenAI, "GPT-4 Technical Report."
[10] Alimardani, "GenAI and the Mirage of Personalised Learning for All."
[11] See Alimardani, "Borderline Disaster"; Yuvaraj, "The Verification-Value Paradox."

a means of assuring that graduates can read, reason, verify, and apply law independently of a machine.[12] Part should deliberately incorporate AI and assess the student's contribution to the human–AI collaboration, the prompts chosen, the sources checked, the errors caught, the reasoning improved, and the judgement exercised. This would make the process, and not only the product, the object of assessment.[13] A third, 'relay' model should combine the two, requiring students either to critique and improve AI-generated work without AI access or to use AI to strengthen work first produced independently. The aim is neither to stop students using GenAI nor to license its unconstrained use. It is to ensure that graduates can outperform, supervise, and improve AI-generated legal work, a capability that, on the evidence, cannot be developed through one-off training but must be built across the degree.[14]

The remainder of the article sets out the methodology and reports models' overall performance against the student cohorts and the 2023 baseline. Variation by question type and by model is examined, as well as the quality of critical analysis and the citation of legal authority. Hallucination and the verification it demands are then considered and the article concludes with implications for the design of law school assessment.

2. Methodology and Limitations

The methodology encompasses multiple stages: exam design and delivery, GenAI prompt and context engineering, response generation, response transcription, and blinded tutor marking, supplemented by non-blinded coordinator marking. Ethical considerations were integrated throughout the research process, including the use of covert data collection and post hoc marker consent.

2.1 Exam Context and Structure

The exams for Criminal Law and Tort Law were summative, in-person, open-book assessments conducted under invigilated conditions. Students were prohibited from using digital devices. However, they were allowed to bring printed materials. The exam duration was three hours and responses were handwritten in the majority of cases (see below).

Criminal Law topics include murder, manslaughter, various forms of assault, and the exercise of police powers. The Criminal Law exam tested students across three questions, cumulatively worth 100 marks and contributing 60% of the final course grade. The distribution was as follows: Question 1, 50 marks, was a complex, multi-issue hypothetical scenario question. Questions 2 and 3, worth 30 and 20 marks, respectively, were structured essay questions requiring critical engagement with legal principles and case law. As Criminal Law relies heavily on both legislation and case law, relevant legislative extracts were supplied in the exam paper; however, students were expected to bring their own case law notes.

Tort Law focuses on liability under negligence and nine intentional torts, including assault, battery, false imprisonment, trespass, and nuisance. The Tort Law exam comprised four questions, also worth 60% of the final grade. To facilitate cross-subject comparison, these scores were normalised to a 100-point scale.[15] The questions were structured as follows: Questions 1 to 3 were worth approximately 23, 33, and 30 marks and consisted of hypothetical problem scenarios engaging students in legal analysis across a variety of Tort Law issues. Question 4 was worth approximately 13 marks and was a short-essay format question requiring analytical commentary. Unlike Criminal Law, the Tort Law exam did not include any legislative excerpts. Students were responsible for bringing all necessary materials, including legislation and case law.

Both subjects comprised essay-style and hypothetical scenario questions. The weight on the latter is usually much greater since problem scenarios are pedagogically central in foundational law and require students to identify issues, state and interpret applicable rules, and apply doctrine to novel fact patterns under time constraints. The inclusion of multiple, layered issues in a single scenario is intended to probe the depth of legal analysis rather than surface recall.

2.2 Prompt and Context Engineering

Two primary approaches to improve a GenAI output are prompt engineering and context engineering. Prompt engineering is crafting instructions and constraints (role framing, audience, step-wise reasoning requests, rubric alignment, length guidance, tone/style, and formatting) to elicit higher-quality answers from a model's pre-trained knowledge.[16] Context engineering is

[12] Lodge, Assessment Reform for the Age of Artificial Intelligence; Lindkvist, "Choosing Not to Use Generative AI in Higher Education."
[13] Winstone, "Black Box Assessment"; Head, "Assessing Law Students in a GenAI World to Create Knowledgeable Future Lawyers."
[14] Alimardani, "Borderline Disaster"; Landy, "Early PLT Student Perceptions of the Integration of Generative Artificial Intelligence in Legal Education"; Lodge, Assessment Reform for the Age of Artificial Intelligence.
[15] Because these figures are rounded, they do not sum precisely to 100.
[16] Alimardani, "Generative Artificial Intelligence vs. Law Students"; Choi, "AI Assistance in Legal Analysis."

supplying curated source material to the model (e.g., statutes, cases, lecture notes, slides, or textbook extracts via the context window or retrieval) and designing how that material is chunked, cited, and grounded.[17]

Our main assumption, reflecting likely students' and legal professionals' practice, is that users will invest in prompt engineering but will not systematically compile and inject bespoke relevant legal materials for each question because this would be time consuming. Accordingly, we provided the prompt and exam questions without any subject-specific sources (i.e., no context engineering). This also meant that our approach focused on the baseline legal reasoning capacity of GenAI. This approach allowed for a more conservative and arguably more robust measure of GenAI's independent analytical ability.[18] That said, as we will explain, all models had access to the internet.

Each exam question was input into GenAI systems separately, rather than as an entire exam paper. This modular generation strategy reduced prompt complexity and encouraged more precise responses by narrowing the model's focus. This approach, however, impacted the output size of the models. More specifically, while a uniform base prompt was initially applied across GenAI platforms for each subject, we observed substantial variability in output length. Some models produced responses so lengthy, exceeding 4000 words, that it became implausible for a student to replicate them by handwriting within a three-hour timeframe. To maintain ecological validity, we attempted to retain the same core prompt across models, adjusting the instructions as necessary to reduce length, so that outputs approximated what a student could plausibly produce under exam conditions. This meant that we needed to trim the prompt, omitting terms such as 'detailed' and 'comprehensive' to encourage GenAI models to generate shorter outputs.

While we attempted to use a similar prompt for both subjects, the prompts for Criminal Law and Tort Law differed mainly to make the output consistent with the expected answers from students. For instance, in Criminal Law at the University of Wollongong School of Law, students are encouraged to use the ILAC (Issue, Law, Application, and Conclusion) method when answering problem-based scenario questions; however, this is not the case for Tort Law.

2.3 GenAI Models and Features

We used nine models across five leading GenAI providers: OpenAI, Anthropic, Google, Perplexity, and xAI. Each model featured unique capabilities and configurations (Table 1). While the GenAI models themselves are a constant variable, there are features that can be activated to enhance the model's output. We briefly outline four variables that were used in this study:

- All models had real-time internet access or conditional web-browsing capabilities, i.e., based on the inquiry (prompt), the model decides whether it needs to browse the internet to provide an accurate answer. This enhances the model's accuracy, as it can access up-to-date information or content not included in its training dataset.
- Some models support enhanced chain-of-thought or planning features that aim to structure analysis more deliberately. Certain models are either inherently designed for high-level reasoning or could be enhanced through configuration to engage in structured argument development (extended thinking).
- We gave some models 'Deep Research' capability.[19] This feature enables the model to conduct a more comprehensive search than an 'internet search' and to aggregate information from numerous sources (e.g., web pages) and compile and output results based on the collected information. This process can take considerably more time than the usual prompting (e.g., 30 minutes vs. 2 minutes).
- Some providers offered preset writing styles (e.g., 'concise') which we found produced more focused, exam-appropriate responses.

In June 2025, across both subjects, we generated a total of 18 exam papers, nine per subject, each representing a combination of model and features. While more AI-generated papers using other GenAI models and features could have made the study more comprehensive, the sheer time-consuming process of such empirical studies rendered this impractical for the current project (see below, 'Ethical Considerations').

[17] Alimardani, "Generative Artificial Intelligence vs. Law Students"; Choi, "AI Assistance in Legal Analysis."
[18] Alimardani, "Generative Artificial Intelligence vs. Law Students"; Choi, "AI Assistance in Legal Analysis."
[19] We use 'Deep Research' as an umbrella term for the advanced search-and-retrieve modes of these three models; xAI brands its version 'Deep Search' and 'Deeper Search' and we used the latter.

Table 1. Details on the 18 AI-generated papers

Paper number	Marking format	Grade /100	Model	Internet access	Extended thinking	Deep Research	Writing style
Crim 1/ Tort 1	Blind paper/blind paper	85/83	Claude Opus 4 (Anthropic)	Yes	Yes	No	Concise
Crim 2/ Tort 2	Non-blind paper/blind paper	75/92	O3 (OpenAI)	Yes (conditional browsing)	Yes	No	N/A
Crim 3/ Tort 3	Non-blind paper/non-blind paper	70/60	GPT-4o (OpenAI)	Yes (conditional browsing)	No	No	N/A
Crim 4/ Tort 4	Blind paper/non-blind paper	79/62	Gemini-2.5-pro (via AI Studio) (Google) ('Gemini')	Yes	Yes	No	N/A
Crim 5/ Tort 5	Non-blind paper/blind paper	61/53	Grok 3 (xAI)	Yes	Yes	No	Concise
Crim 6/ Tort 6	Blind paper/blind paper	80/48	Perplexity Unbiased Reasoning Model R1 1776 ('Perplexity Reasoning')	Yes	Yes	No	N/A
Crim 7/ Tort 7	Blind paper (printed)/blind paper (printed)	84/53	Perplexity (Deep Research)	Yes	Yes	Yes	N/A
Crim 8/ Tort 8	Blind paper (printed)/non-blind paper (printed)	73/73	GPT-4.5 (Deep Research) (OpenAI)	Yes	Yes	Yes	N/A
Crim 9/ Tort 9	Blind paper (printed)/blind paper (printed)	80/70	Grok 3 (Deeper Research) (xAI)	Yes	Yes	Yes	Concise

2.4 Exam Papers Transcription and Marking Process

From the nine papers per subject, six were handwritten verbatim by the subject coordinators and three were printed, simulating the reasonable adjustment provisions available for students with specific needs. We removed or neutralised elements that would be anachronistic in an exam paper. For instance, we excluded direct hyperlinks that would not appear in typical student responses. However, we preserved everything else, including any erroneous references the models produced.

AI papers were mixed with student exam papers and allocated to tutors for marking under normal processes. Tutors were not told that some papers were AI-generated. After marking, tutors were debriefed and asked for retrospective consent to include their marks in the study; four of six consented. Grades for the papers lacking tutor consent were discarded by the lead author, and digital versions of those AI papers (three for each subject) were sent to the subject coordinators for marking. As such, while subject coordinators, as authors of this study, were aware of the nature of AI-generated papers, they marked them independently and were unaware of their original grade.

2.5 Ethical Considerations

The study underwent institutional ethics review and received ethical approval from the Social Sciences Human Research Ethics Committee of the University of Wollongong, Reference 2025/107.

A covert method was used to mitigate observer effects that would have compromised the study's validity. Prior notification to tutors would likely have led to changes in marking behaviours, such as increased vigilance for AI-generated content or unconscious bias, thereby distorting genuine grading. Conducting covert empirical research in educational settings nonetheless raises ethical concerns. While the research posed no tangible risk to participants and arguably mirrored real-world conditions in which tutors may already unknowingly mark GenAI-assisted student work in non-invigilated assessments, it risked undermining professional relations. Tutors may have perceived the covert inclusion of AI-generated papers as a misuse of their intellectual labour or a breach of academic collegiality. Tutors may also have felt, in hindsight, that they had been unwitting subjects of an AI detection competency test they were never told they were taking. A further practical concern for such projects

is considerable time investment. After the generation of AI answers, handwriting transcription, and marking, the resulting data could be rendered unusable if post hoc consent were withheld by all or most participants.

Taken together, these considerations illustrate the logistical and ethical complexity inherent in covert empirical research and in studying the impact of AI technologies on education more broadly.

2.6 Limitations

This study is subject to several limitations that must be acknowledged when interpreting the findings. The effectiveness of GenAI responses is directly influenced by the quality of prompt engineering. While we endeavoured to standardise prompts across models, some variation was necessary to control output length and ensure exam realism. These modifications may have unintentionally favoured certain models, introducing a potential confounding variable. Furthermore, our design did not allow for a comparative analysis of how prompt variation affects performance, a valuable avenue for future research. We do not believe this limitation had a significant impact on this study given that the models are constantly being updated and the same prompt may generate different outputs. A related limitation is that each paper comprised only one generated response per question. Because GenAI output is stochastic, each paper-level grade and corresponding percentile rank therefore reflects a single realised set of outputs rather than a precise estimate of a model's expected performance. Future research should use repeated generations for each model question combination to quantify within-model variability and improve the precision of these estimates. Our results should be interpreted as an estimate of what a capable user could achieve with modest prompt skill, not a ceiling. Further, results may not be generalisable across all areas of legal education. Topic selection and question format may affect GenAI performance in other law subjects such as Contract Law, Constitutional Law, or Equity.

Despite rigorous quality assurance processes within the School of Law, including moderation and peer review of exams, marking in legal education remains inherently subjective. As a result, some GenAI responses may have been marked more harshly or favourably than comparable student submissions. Our protocol assesses the ability of off-the-shelf GenAI to produce exam-adequate legal analysis under generous generation conditions (no fatigue, perfect legibility, no time decay in handwriting speed). This is not equivalent to student mastery of law, nor does it model within-exam constraints (anxiety, triage, drafting discipline). Further, we generated each question separately to reduce prompt length and improve focus. A human candidate must allocate time across questions, hold earlier analysis in working memory, and manage cognitive switching costs. Our approach may modestly favour AI by removing inter-question time trade-offs. The finding that AI can match or exceed typical student performance speaks to assessment vulnerability, not to the equivalence of professional competence.

3. Analysis and Findings

This section presents the empirical findings of the 2025 study and situates them against the 2023 baseline. The analysis begins with overall performance, comparing AI-generated papers with student cohorts across the two subjects, question types, and individual models. It then turns to the qualitative skills most valued in legal education: critical analysis, structured reasoning, and citation of primary authority. The final section examines the hallucination rate among AI models, which is one of the most significant limitations of GenAI.

3.1 Preliminary Findings

The 2025 analysis drew on anonymised grades from 302 students in criminal law and 273 students in tort law to benchmark the performance of AI-generated papers. The results represent a marked improvement over the 2023 baseline, where the average percentile rank was 22.1 and the top three AI papers outperformed 86.1%, 75.0%, and 25.8% of students, respectively.[20] In contrast, the 2025 cohort of AI-generated papers across both subjects achieved a mean percentile rank of approximately 72, with seven of 18 papers ranking at or above the 90th percentile (see Table 2).

Direct comparison between the two studies must be approached with caution due to methodological differences. The 2025 study used different prompting strategies and did not provide any subject-specific legal content (such as lecture transcriptions or content summaries) to any of the models, whereas such materials were supplied to some models in 2023. This distinction makes the performance gains even more striking: despite receiving less context-specific input, the 2025 models substantially outperformed their 2023 counterparts. This trajectory is consistent with broader advances in Deep Research capabilities and chain-of-thought reasoning observed across major language models over the two-year intervening period.

[20] Alimardani, "Generative Artificial Intelligence vs. Law Students."

While raw grades are not directly comparable between Criminal Law and Tort Law due to differences in the difficulty of the subjects for both AI and human test-takers, percentile rankings in Table 2 provide a more robust basis for evaluating relative performance.

As seen in Table 2, in Criminal Law, five of the nine AI-generated papers outperformed 90% of students, with the top performer (Claude Opus 4, P1) exceeding 97% of the cohort. Performance in Tort Law was more modest overall, although two papers still surpassed the 90th percentile, with the highest (O3, P2) reaching the 99th percentile, the single strongest result across both subjects. On average, models outperformed 82.5% of students in Criminal Law and 61.0% in Tort Law. It is worth noting that Criminal Law is a first-year subject, whereas Tort Law is usually taken in the third year. The lack of experience in writing law exams may partially explain why there was a much larger gap between the models' and students' performance in Criminal Law compared to Tort Law.

A notable finding is that despite OpenAI's widely cited 2023 claim that GPT-4 outperformed 90% of US bar exam takers, only one of the six papers generated by three OpenAI models across the two subjects exceeded the 90th percentile.[21] While the bar exam and Australian law school assessments differ in both format and jurisdiction, the present findings echo concerns raised in the 2023 study, particularly regarding the impact of such claims on our understanding of AI capabilities in law.

Table 2. Performance of AI models out of 100 and against students

		2025 AI individual papers /100		**2025 Percentile ranking (AI better than % of students)**	
Paper number	**AI model used**	**Criminal Law**	**Tort Law**	**Criminal Law**	**Tort Law**
P1	Claude Opus 4	85	83	97.0%	90.0%
P2	O3 (OpenAI)	75*	92	80.7%*	99.0%
P3	GPT-4o (OpenAI)	70*	60*	68.8%*	51.6%*
P4	Gemini (via AI Studio)	79	62*	92.7%	57.9%*
P5	Grok 3	61*	53	42.5%*	37.9%
P6	Perplexity Reasoning	80	48	93.0%	28.4%
P7	Perplexity (Deep Research)	84	53	96.3%	37.9%
P8	GPT-4.5 (Deep Research) (OpenAI)	73	73*	78.1%	75.3%*
P9	Grok 3 (Deeper Research)	80	70	93.0%	70.7%
Average		**76%**	**66%**	**82.5%**	**61.0%**

* Papers that were marked by coordinators (two of the authors).

Model performance was not uniform across the two subjects, and the degree of variation itself is analytically significant. Some models demonstrated remarkable consistency: Claude Opus 4 (P1), for instance, achieved scores of 85 in Criminal Law and 83 in Tort Law, outperforming 97% and 90% of students, respectively. Other models, however, exhibited striking subject-dependent disparities. Perplexity's reasoning model (P6) and Perplexity Deep Research (P7) both achieved distinctions in Criminal Law, exceeding the 90th percentile. Yet in Tort Law, the same models scored only 48 and 53, placing them in the 28th and 38th percentiles, respectively, over 50 percentile points lower. This inconsistency suggests that current GenAI capabilities are not evenly distributed across legal domains.

In the 2023 study, AI models performed markedly better on essay-style questions than on problem-based scenarios.[22] The 2025 results tell a different story. In Criminal Law, models performed almost identically across both question types, averaging 76.0% on the problem question (Question 1) and 76.8% on essay questions (Questions 2 and 3) (Table 3). This convergence is significant: it suggests that the earlier weakness in applied legal reasoning have narrowed substantially. In both question types, models outperformed students, by margins of 16.4 and 13.2 percentage points, respectively.

[21] OpenAI, "GPT-4 Technical Report."
[22] Alimardani, "Generative Artificial Intelligence vs. Law Students."

In Tort Law, model performance on problem-based questions (Questions 1–3) was relatively consistent, with averages ranging from 62.2% to 65.5%. On the essay question, however, models averaged 84.0%, a substantial uplift. The comparison with student performance is particularly revealing: on problem questions, the AI advantage was marginal (less than one percentage point), suggesting that students and models are performing at near-parity on applied Tort analysis (Table 4). On the essay question, however, models outperformed students by 43.2 percentage points. This large gap likely reflects the essay's reliance on broad doctrinal knowledge and structured argumentation, areas where GenAI's capacity for synthesising large bodies of legal content gives it a pronounced advantage.

Table 3. Comparison between question types average in Criminal Law (grade out of 100%)

Criminal Law	AI %	Students %
Problem question (Q1)	76.0%	59.6%
Essay questions average (Q2 & Q3)	76.8%	63.6%

Table 4. Comparison between question types average in Tort Law (grade out of 100%)

Tort Law	AI %	Students %
Problem questions average (Q1–Q3)	63.3%	62.5%
Essay question (Q4)	84.0%	40.8%

3.2 Critical Analysis

Critical analysis is one of the most valued competencies in legal education. It requires structured and systematic analysis of the content, evaluation of various options, and identification of what is important in a given context.[23] This is a skill that students often find difficult to develop. Critical and creative thinking have traditionally been regarded as distinctively human capabilities and as potential areas of comparative advantage over AI.[24]

To assess the capability of AI models in this area, we adopted a benchmarking approach: a score of 5 out of 10 was assigned as the expected level for a typical student, and 10 as the standard for an excellent student. Each model's critical analysis was then evaluated relative to these benchmarks. The same method was used in other sections below, i.e., 'Citing Legal Authority,' and 'Source Types and Reliability.' The 1–10 benchmark ratings were assigned by the relevant subject coordinator, based on their assessment of each GenAI response against the relevant student benchmark. The responses were not independently double-rated and accordingly, inter-rater reliability could not be calculated.

In Criminal Law, all models performed at or above the level of a typical student (Table 5). The standout performers were Perplexity Deep Research (P7), followed by Gemini (P4) and Claude Opus 4 (P1), with scores that approached or met the 'excellent student' threshold.

Table 5. Criminal Law — critical analysis (compared to average and excellent students)

Paper No.	P1	P2	P3	P4	P5	P6	P7	P8	P9
Question/Model	Claude Opus 4	O3	GPT-4o	Gemini (via AI Studio)	Grok 3	Perplexity Reasoning	Perplexity (Deep Research)	GPT-4.5 (Deep Research)	Grok 3 (Deeper Research)
Q1	9	6	7	10	8	7	9	7	8
Q2	7	9	7	8	5	6	10	10	5
Q3	8	9	8	7	5	8	9	6	9
Avg.	8	8	7.33	8.33	6	7	9.33	7.67	7.33

[23] The Law Society of New South Wales, *Solicitor Capability Framework* (2025), https://www.lawsociety.com.au/sites/default/files/2025-08/LS4645_CPD_CapabilityFramework_2025-08-29.pdf.
[24] Mawn, "Using Artificial Intelligence as a Mindtool"; Marr, "6 Skills You Need to Stand Out in the Age of AI."

Table 6. Tort Law — critical analysis (compared to average and excellent students)

Paper No.	P1	P2	P3	P4	P5	P6	P7	P8	P9
Question/Model	Claude Opus 4	O3	GPT-4o	Gemini (via AI Studio)	Grok 3	Perplexity Reasoning	Perplexity (Deep Research)	GPT-4.5 (Deep Research)	Grok 3 (Deeper Research)
Q1	7	6	5	4	7	2	6	10	4
Q2	8	5	7	6	8	3	6	8	6
Q3	5	4	4	7	7	4	8	9	7
Q4	9	10	9	10	10	7	7	7	5
Avg.	7.25	6.25	6.25	6.75	8	4	6.75	8.5	5.5

In Tort Law, on average, all models except Perplexity Reasoning (P6, scoring 4.0) outperformed the typical student benchmark. Five models scored below 5 on at least one of the four questions (Table 6), revealing that critical analysis remains uneven across individual questions even where overall averages are strong. GPT-4.5 with Deep Research (P8) was the strongest performer, scoring 8.5.

These results provide evidence of a considerable advance over the 2023 study, in which critical analysis was identified as one of the most significant weaknesses of GenAI, a limitation specifically noted in the interviews by the tutors who blind marked those papers.[25] One possible contributor to the observed improvement is the emergence of models with 'thinking' capabilities, i.e., systems that employ an explicit chain-of-thought reasoning process before generating a response (extended thinking). However, this study did not isolate the effect of those capabilities from differences in model generation, prompting, internet access, or assessment conditions. This inner reasoning allows models to consider information iteratively, weigh competing interpretations, and revise their positions, in ways that closely approximate human analytical reasoning.[26]

Examination of the models' chain-of-thought outputs reveals how this process operates in practice. Several models demonstrated the capacity to simulate the kind of adversarial reasoning central to legal analysis, constructing arguments, anticipating counterarguments, and weighing evidence. In the Tort Law exam, Perplexity Reasoning (P6) produced the following chain-of-thought passage (notes in square brackets are added by authors): [27]

> For Dr Badlukko, even though the surgery was done correctly, the failure to warn about the risk is the breach. So the injury from the complication is directly linked to the lack of warning. Eugene would argue that had he known, he might not have opted for surgery, thus avoiding the complication. But he was eager to heal quickly, so maybe he would have taken the risk anyway [consideration of how a party might have reasoned given circumstances]. However [adding a different perspective by referring to a legal authority], under Rogers, the test is whether a reasonable patient would consider the risk significant. Even a 0.001% risk might be material if it's a severe outcome like permanent hearing loss, especially given Eugene's existing hearing issues [analysing what would be considered significant risk for a specific person given their conditions]. The survey of doctors (75% would disclose) supports that the risk was considered material by peers, so the doctor's failure breaches the standard [using evidence provided in the problem scenario to support an argument].

This passage is analytically sophisticated. The model constructs the plaintiff's argument, immediately identifies a factual weakness (Eugene's eagerness to heal), pivots to the legal test under the authoritative case of *Rogers v Whitaker*,[28] and applies it to the specific circumstances of the claimant. It then retrieves relevant content from the problem scenario to support its

[25] Alimardani, "Generative Artificial Intelligence vs. Law Students."
[26] Jaech, Aaron, Adam Kalai, Adam Lerer et al., OpenAI O1 System Card.
[27] This was a hypothetical scenario (around 450 words) where students had under 60 minutes to write advice to the plaintiff (Eugene) as to whether two other parties were liable to him in negligence (one of whom is Dr Badlukko). After an initial injury, arguably due to the negligence of the first defendant, Eugene was operated on by Dr Badlukko. He was presented with two options: do nothing other than observe the natural healing process, which typically takes two years, or undergo surgery, with full healing expected within a couple of weeks. Eugene selected the latter option, yet Dr Badlukko had not informed him of the inherent risk of this surgical procedure which included a well-known, but exceedingly rare, complication. This complication eventuated, and it was especially impactful for Eugene because of an existing, known condition. The core issue students faced was determining if Dr Badlukko had fallen below the required standard of care by not informing Eugene of a specific risk in circumstances where other doctors would not have advised of the risk and the patient was keen to heal quickly.
[28] *Rogers v Whitaker* (1992) 175 CLR 479.

conclusion. This kind of multilayered, evidence-grounded reasoning is precisely what law examiners reward, and what many students struggle to produce under exam conditions.

Another striking feature of thinking models is their capacity for self-correction, a process that mirrors the iterative reasoning that legal educators seek to cultivate in students. Grok 3 (P5) provided a vivid illustration in Criminal Law (notes in square brackets are added by authors): [29]

> So, for murder, I need to look at whether Barry had the intent to kill Ian, or at least the intent to cause him grievous bodily harm, or if he was recklessly indifferent to human life [reiterating the question].
>
> Let me think about the intent to kill first. Did Barry intend to kill Ian? From the facts, it seems like Barry was trying to stab Ian in the knee, not somewhere that would typically be fatal. He even said to Nina that he only meant to stab Ian in the knee to get him to move. So, it doesn't sound like he wanted to kill Ian. But intent can be a bit tricky; it's about what was in his mind at the time [an accurate analysis of difficulty to prove intent].
>
> Wait, actually, in criminal law, intent to kill doesn't necessarily mean he wanted Ian to die [correct]. It's more about whether he foresaw death as a probable consequence of his actions [incorrect, as that is only one of the heads of murder]. But no, that's not quite right. For murder, the intent to kill is a specific intent; he has to have intended to cause death [the model corrected itself]. Alternatively, if he intended to cause grievous bodily harm, and death resulted, that can also be murder [correct].
>
> So, perhaps I should consider whether Barry intended to cause grievous bodily harm to Ian. Grievous bodily harm, or GBH, is defined in NSW law as really serious harm. Stabbing someone in the knee might not necessarily be seen as intending to cause really serious harm [controversial assessment], but then again, using a screwdriver to stab someone could be argued as intending to cause serious injury [more accurate assessment].

What is remarkable here is the visible process of error and correction. The model initially misstates the legal test for intent, recognises the error, and revises its position, arriving at the correct formulation. The same approach occurred with the grievous bodily harm (GBH) analysis in the final paragraph, where the model updated its assessment of circumstances to one that was more reasonable.

This mirrors the kind of iterative reasoning that is more or less 'human-like': testing a proposition, identifying its flaws, and refining the analysis. It also, however, raises a deeper question about the nature of this process. As another passage from the same model illustrates, 'Case law has held that GBH means really serious bodily injury, as per *R v Perks* [1986] or something like that. I think it's *R v Brown* (1994) actually, but I need to recall properly. Anyway, the point is …' The sentence mimics the *experience* of uncertain recall, despite AI models having no genuine memory or intuition to draw upon.

3.3 Citing Legal Authority: Case Law and Legislation

The citation of authoritative legal sources, principally case law and legislation, is a foundational expectation in legal writing. Accurate citation serves multiple functions: it demonstrates the writer's command of relevant doctrine, anchors analytical arguments in recognised authority, and enables the reader to verify the provided propositions. For these reasons, the quality and accuracy of legal citations produced by GenAI models is a critical metric in evaluating their suitability for law assessment tasks.

For the case law citations, in Criminal Law, Google Gemini (P4) produced the highest score (out of 10 relative to an excellent student), with Claude Opus 4 (P1) ranking second. However, Claude Opus 4 also exhibited the second highest rates of hallucinated case citations. This alarming finding indicates that models that cite more cases may also generate more fictitious cases, a pattern that complicates any straightforward assessment of citation quality. In Tort Law, most models demonstrated adequate case law citation performance, with the notable exception of Perplexity Reasoning (P6). An unexpected result was that standard Grok 3 (P5) outperformed its Deeper Research-enabled counterpart (P9), a finding discussed further below.

With respect to legislation citations, in Criminal Law, models achieved the highest overall scores for legislative citations compared to case law. However, performance varied considerably across providers, and strong case law citation did not predict

[29] In this hypothetical scenario, students were asked, inter alia, to advise whether the charge of murder could be established against Barry arising out of a mêlée in a car park. Students were specifically instructed to explain the elements of each head of murder, providing relevant authority, and, by applying the law to the facts, come to a conclusion as to the likelihood of success on each basis/head of murder.

strong legislative citation. In Tort Law, models performed notably worse on legislative citations than on case law, which is the opposite of the Criminal Law pattern.

Contrary to expectations, access to Deep Research capabilities (P7, P8, P9) did not confer a clear advantage in the frequency of citing relevant case law, with citation rates showing no meaningful improvement over standard models. In Criminal Law, GPT-4.5 with Deep Research (P8) missed the mark on Question 1, providing only one case citation. Review of its research process — which, like the thinking process, is not included in the final output — revealed that the model had identified three relevant cases during its search phase but failed to cite any of them in its response.

This discrepancy between research and output was not unique to GPT-4.5. Several models referred to cases in their chain-of-thought reasoning that did not appear in their final answers. In some instances, the model appeared to have assessed the case as less relevant than alternatives; in others, the omission was less explicable. The O3 model (P2), for example, explicitly noted in its reasoning, 'When discussing the bail application process, I'll use cases like *DPP (NSW) v Tikomaimaleya* [2015], *R v Zayad*, and possibly *R v Alqudsi* [2015] NSWSC 1229. While *DPP (Cth) v Barbaro* is relevant for sentencing, I may focus on the more recent *R v Magdalene* [2024] NSWSC 1123 for bail-related context.' Yet only one of these cases, *DPP (NSW) v Tikomaimaleya*, which was the most relevant in answering the exam question, appeared in the final output.

A further complication arose from the citation format adopted by several models, particularly those with Deep Research capabilities. These models appeared optimised to cite sources in hyperlinked format, either as numbered reference links, hyperlinked text, or raw URLs. While this format may be appropriate in other disciplines or professional contexts, legal writing requires citation of primary authorities, case names, and legislation within the text itself. As outlined in the 'Methodology and Limitations' section, hyperlinks and links were removed from the AI-generated responses prior to marking and therefore those citations were not credited by markers.

This formatting tendency appears connected to a deeper issue with how Deep Research models engage with source material. These models conduct an explicit search-and-retrieval phase before generating their response, and in doing so, they predominantly access secondary sources such as law firm blog posts, practitioner summaries, and general commentary, where primary authorities are not cited in full. As a result, the models' outputs tend to reference these secondary sources via hyperlinks rather than citing the underlying case law and legislation directly. Unlike other models, Deep Research can reduce reliance on a model's internal knowledge (i.e., training dataset) by retrieving external sources and passing them to the model as context. Therefore, while this may reduce the possibility of hallucination, it may also result in not directly referencing primary sources.

3.4 Source Types and Reliability

Across both subjects, we identified four categories of internet sources upon which models drew. The first category was primary legislative sources, including the NSW Legislation website, NSW Parliament website, AustLII, and JADE. This is clearly illustrated in various papers, such as Perplexity Deep Research (P7) answer, which referenced 'the flow charts contained within section 16 of the Act' and provided a direct link to the *Bail Act* on the NSW Legislation website. Second, legal principles drawn from unreliable secondary sources, including Wikipedia and law firm blog posts. In our experience, these sources mostly provide only generic references to legislative sections and rarely cite relevant case law. Some models also cited platforms such as StudentVIP that aggregate student-uploaded notes and lecturer materials. Third, secondary sources such as the bench books published by the Judicial Commission of New South Wales that are reliable legal sources. Fourth, academic literature accessed either directly through open-access journals, including the *Alternative Law Journal*, *University of Queensland Law Journal*, *UNSW Law Journal*, *Journal of Legal Medicine*, and *International Psychogeriatrics*, or indirectly through repositories such as AustLII, Semantic Scholar, and PubMed. As we will discuss below, we could identify only one instance of a direct link to a case law database.

The breadth of sources accessed by models in this study is notable and represents an advance over the 2023 findings.[30] However, we could not find any indication that the models distinguish between sources of differing authority and reliability: a StudentVIP summary and a Judicial Commission bench book appear to be treated with equal weight. This indiscriminate sourcing exemplifies the broader problem of 'AI slop,' i.e., the tendency of generative models to absorb and reproduce content without regard to its provenance, reliability, or authority.[31] In a legal context, where the hierarchy of sources carries normative weight, this failing is consequential.

[30] Alimardani, "Generative Artificial Intelligence vs. Law Students."
[31] Kommers, "Why Slop Matters."

Nevertheless, the models were able to demonstrate awareness of the temporal limitations of their training data. O3 (P2) noted in its chain of thought:

> I need to consider whether any recent changes have been made to the Civil Liability Act, especially regarding mental harm. After checking, it seems the 2024 bill primarily focuses on child abuse and doesn't affect the provisions for mental harm. Therefore, the relevant mental harm provisions from the 2002 Act remain unchanged.

This suggests that some models may attempt, at least in some circumstances, to check whether their knowledge remains current. This is notable given that the 2023 study identified reliance on outdated training data as a significant limitation.[32]

3.4.1 Misleading citations

Despite providing direct links to sources which on the surface lend the output an appearance of reliability and thorough research, we identified instances where models cited URLs that did not support their outlined discussions. For example, Perplexity Deep Research (P7), responding to a question concerning trespass under civil law, cited a source that addressed trespass under criminal law. In another section, the same model linked to an online source containing legal information that, while consistent with Australian law, in fact addressed Canadian jurisdiction. Gemini (P4) referred to a relevant legislative provision but incorrectly stated that the section had been repealed. We could not locate this claim from the cited link by Gemini. In the same response, Gemini attributed the phrase 'staggering injustice' to a judge in the context of section 25B of the *Crimes Act 1900* (NSW). Once again, we could not locate such a reference in the cited source. Notably, none of these sources cited by Gemini provided links to specific web pages. They only referenced the domain (e.g., https://legislation.nsw.gov.au), making verification more difficult and the appearance of authority more misleading.

Whether this pattern reflects straightforward hallucination (which we discuss below under 'Hallucination') or a more systematic tendency is an open question. At a design level, however, it raises the concern that models may be optimised to produce citation-like outputs, attaching plausibly relevant URLs to generated propositions, regardless of whether the cited source supports the claim. The effect is that the reader is presented with an output that appears well-sourced but is in key respects fabricated.

Another notable anomaly concerned the JADE website (Jade.io). Perplexity Deep Research (P7) cited a case via a direct Jade.io link, the only instance across all models of a direct link to a case law database. This is surprising because unlike AustLII (which is freely accessible), Jade.io requires users to provide an email address and complete a verification step before granting access. Since no computer-use capabilities (i.e., where models operate a browser interface, take screenshots, and navigate pages as a human would) or email plug-ins (i.e., giving the AI models access to an email account to activate the verification link) were enabled for any model in this study, it is unclear how Perplexity accessed this resource. Two possibilities present themselves: either Perplexity has a commercial arrangement with Jade.io or the model encountered the link on another website and reproduced it without accessing the database, providing an illusion of direct access to primary legal materials.

3.4.2 Gaps in citation coverage

Across both subjects, we identified key cases and legislative provisions that no model cited, despite their centrality to the topics assessed. These authorities were discussed in lectures, tutorials, and the prescribed textbook, materials that were not provided to any model in this study. Their absence from all AI-generated responses demonstrates the importance of context engineering: without access to subject-specific teaching materials, models rely on any publicly accessible legal content, which may not align with the authorities a particular course emphasises. That said, it remains unclear whether models would necessarily cite these authorities even if provided with the relevant materials and citations.

A related barrier is that some legal databases and websites actively block AI web crawling. This may limit the extent to which AI systems can access or incorporate some legal materials, even where those materials are otherwise discoverable through search.[33] Even where access is possible, models may rely on their training data rather than conducting a fresh internet search or may, as some researchers have suggested, be optimised for efficiency in ways that lead them to truncate their search process.[34] Models used in this study do not have the capacity to access legislative material that requires multiple navigation steps and is not indexed by standard search engines. Computer-use capabilities could in principle address this limitation, but at the time of writing this article, this approach is slow, expensive, and inaccurate.[35]

[32] Alimardani, "Generative Artificial Intelligence vs. Law Students."
[33] See Steinacker-Olsztyn, "Is Misinformation More Open?"
[34] Sun, "AutoSearch."
[35] See Abhyankar, "Osworld-Human."

Overall, legal research requires skill and knowledge to locate and identify relevant and reliable sources of law. However, many AI models failed to demonstrate this ability. In other words, the models do not appear to have the legal judgment needed to determine which sources are relevant and reliable for legal argument. It is expected that companies offering AI tools with access to legal databases (e.g., LexisNexis and Thomson Reuters) will perform better at drawing on relevant sources. Performance within these databases also depends heavily on AI models' ability to retrieve pertinent information from large volumes of data.[36] A 2025 empirical study assessing the capabilities of these specialised models found that, while they outperformed the general AI models (such as those examined in this study), they nonetheless have notable limitations.[37] Such evaluations should be repeated in the coming years as AI models continue to advance.

3.5 Hallucination

In this study, a hallucination is defined as a model output that is materially incorrect and demonstrably fabricated, invented, or otherwise unsupported by any relevant source or reasoning process. The definition is deliberately narrow. Outputs grounded in outdated training data, superseded legal materials, or stale information retrieved from external sources were not classified as hallucinations, provided they remained relevant to the legal issue. Similarly, responses containing weak, incomplete, or poor-quality reasoning were excluded from this category where the underlying content could be traced to a relevant authority or factual context. In such cases, the deficiency was better understood as reliance on outdated authority or flawed reasoning rather than fabrication, and was accordingly assessed under metrics concerning identification of relevant laws and legal authorities and critical analysis.

Drawing this distinction is important because hallucination, strictly understood, is a qualitatively different failure mode from error and inaccuracy. An outdated citation may mislead, but it nonetheless points to a real authority that existed and can be corrected. A hallucinated citation constructs an authority that does not exist, foreclosing correction and potentially introducing fabricated propositions into the legal record. The former is a knowledge gap and the latter is a generative artefact, and the two warrant different mitigation strategies.

Across both subjects, models with access to Deep Research exhibited substantially lower hallucination rates than their standard counterparts, a pattern consistent with the hypothesis that explicit search-and-retrieval processes constrain generative fabrication by anchoring outputs to externally verifiable sources. In Criminal Law, the three Deep Research models together produced only one hallucination (see Table 7). In Tort Law, two of the three produced none (see Table 8). Given that hallucination has consistently been identified as one of the most significant limitations of GenAI in legal applications, this represents an important development and suggests that with appropriate context engineering, current GenAI models can achieve a relatively lower rate of fabrication.

This advantage should not, however, be overstated. As the accuracy–coverage analysis below demonstrates, part of (or arguably a significant share of) the reduction in hallucination among Deep Research models is a function of reduced citation volume: a model that cites fewer authorities has fewer opportunities to hallucinate.[38] The overall volume of hallucination was significantly lower in Tort Law than in Criminal Law, despite the fact that Tort Law models generally cited more sources.

Table 7. Criminal Law — hallucination count for case citation and legal content

	Hallucination count								
Question	**Claude Opus 4**	**O3**	**GPT-4o**	**Gemini (via AI Studio)**	**Grok 3**	**Perplexity Reasoning**	**Perplexity (Deep Research)**	**GPT-4.5 (Deep Research)**	**Grok 3 (Deeper Research)**
Q1	4	7	11	2	2	4	0	0	1
Q2	11	1	4	0	4	6	0	0	0
Q3	5	8	2	1	0	0	0	0	0
Total	20	16	17	3	6	10	0	0	1

[36] Gao, "Retrieval-Augmented Generation for Large Language Models."
[37] Magesh, "Hallucination-Free?"
[38] That said, this analysis is most reliable when models are compared within a single subject area (see below).

Table 8. Tort Law — hallucination count in case citation and legal content

	Hallucination count								
Question	**Claude Opus 4**	**O3**	**GPT-4o**	**Gemini (via AI Studio)**	**Grok 3**	**Perplexity Reasoning**	**Perplexity (Deep Research)**	**GPT-4.5 (Deep Research)**	**Grok 3 (Deeper Research)**
Q1	2	4	3	0	0	0	0	0	1
Q2	4	0	0	1	0	2	0	0	1
Q3	0	1	0	0	0	0	0	0	0
Q4	0	0	0	0	0	0	0	0	2
Total	6	5	3	1	0	2	0	0	4

3.5.1 Accuracy and coverage in case law citations

As discussed above, a low hallucination rate is a necessary but not sufficient condition for reliable citation. Data inspection revealed some of the models, particularly those with Deep Research capability, achieved the lowest hallucination counts partly because they cited few authorities and many of their citations took the form of hyperlinks that were removed from the marked output (see 'Methodology and Limitations'). GPT-4.5 with Deep Research (P8), for instance, cited only two cases in Criminal Law, with no hallucinations: a 100% reliability rate atop a very narrow base. Characterising such a result as 'superior performance' without further assessment would be misleading.

To capture this interaction, model performance is evaluated along two dimensions of case law citations. The first, *accuracy*, measures the proportion of cited authorities that were not hallucinated. The second, *coverage*, measures the number of authorities cited relative to the expected number for the question. Because there is no authoritative benchmark for how many cases a well-reasoned response 'ought' to cite in a problem scenario, a working proxy was adopted. The highest number of case citations by any model on a given question was treated as the expected number, and each model's coverage was expressed as a proportion of that ceiling (Table 9 and 10). This proxy has obvious limitations. It defines the expected standard by model performance rather than by disciplinary convention, and it assumes that the most prolific citer reflects a reasonable upper bound rather than over-citation. It is adopted here as an indicative measure only, intended to avoid the distortion that would arise from rewarding models that cite almost nothing, and its results should be read accordingly.

If reliability and coverage are treated as equally weighted, the best-performing models are those that cluster toward the top right of the accuracy–coverage plot. In Criminal Law, Grok 3 (P5) occupies that position, combining relatively high accuracy with relatively high coverage (see Figure 1).

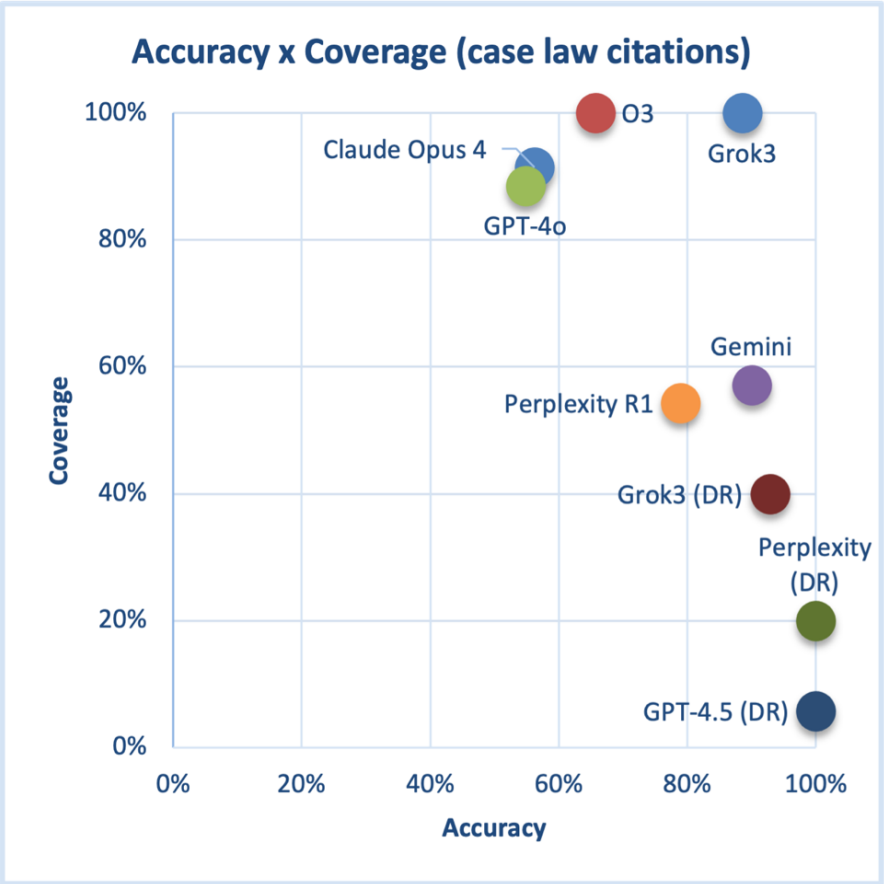

Figure 1. Criminal Law accuracy × coverage (DR stands for Deep Research)

The Deep Research models (indicated by 'DR') are placed toward the bottom-right quadrant, signalling high accuracy but low coverage, while the high-coverage models (Grok 3 excepted), are distributed toward the lower-accuracy region. The shape of this distribution is itself informative. It suggests that, in Criminal Law, there is a practical trade-off between citation volume and citation accuracy that most current models have not resolved.

Table 9. Criminal Law: Case law hallucination divided by number of citations (repeated citations and hallucinations are counted) — lower is better

	Number of hallucinations/number of citations								
Question	**Claude Opus 4**	**O3**	**GPT-4o**	**Gemini (via AI Studio)**	**Grok 3**	**Perplexity Reasoning**	**Perplexity (Deep Research)**	**GPT-4.5 (Deep Research)**	**Grok 3 (Deeper Research)**
Q1	3/17	7/21	8/20	2/17	1/27	4/16	0/2	0/1	1/11
Q2	7/8	1/6	4/6	0/2	3/5	0/3	0/4	0/1	0/1
Q3	4/7	4/8	2/5	0/1	0/3	0/0	0/1	0/0	0/2
Total	14/32 (44%)	12/35 (34%)	14/31 (45%)	2/20 (10%)	4/35 (11%)	4/19 (21%)	0/7 (0%)	0/2 (0%)	1/14 (7%)

In Tort Law, the distribution is notably different. O3 (P2) and Claude Opus 4 (P1) occupy positions close to the top-right corner, combining high accuracy with high coverage (Figure 2). The appearance of these two models as top performers in Tort Law is on its face in tension with the fact that O3 has the highest number of hallucinations (four) and Claude sits in fourth place (two).

The apparent contradiction resolves once the two dimensions are separated: a model may produce a high absolute number of hallucinations while still achieving a favourable reliability ratio if its overall citation volume is high enough. This illustrates the value of disaggregating hallucination counts from accuracy rates, and cautions against interpreting either metric in isolation. More broadly, whereas the Criminal Law plot shows models distributed along a clear accuracy–coverage trade-off, in Tort Law, the distribution leans toward the accuracy axis, with fewer models achieving high coverage.

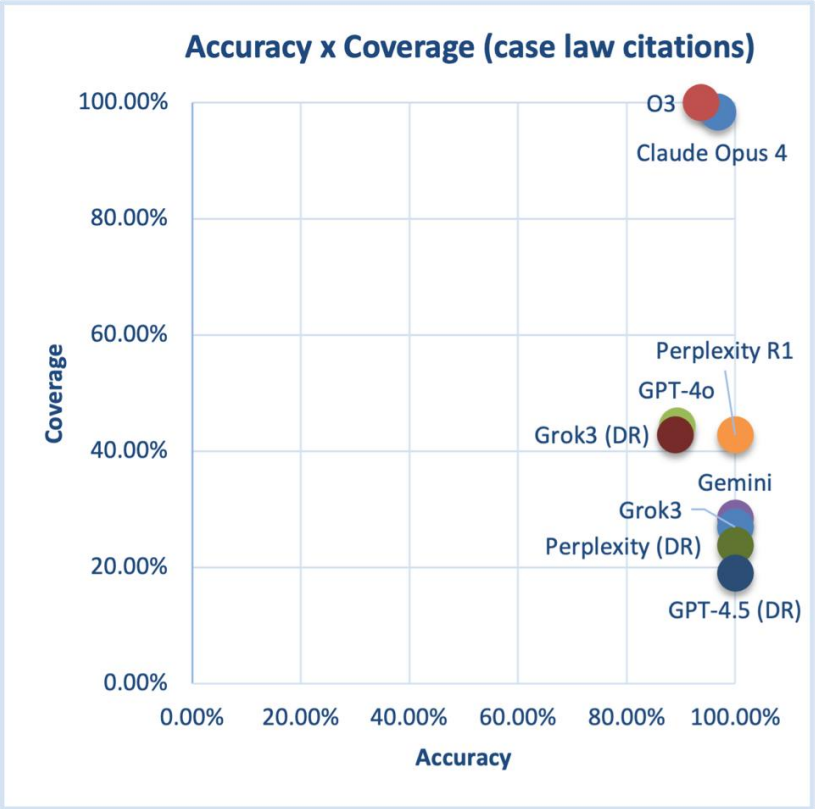

Figure 2. Tort Law accuracy × coverage (DR stands for Deep Research)

A substantial part of the Deep Research models' low coverage is attributable to citation format rather than to any deficiency in legal knowledge retrieval. As discussed above, these models predominantly cite via hyperlinks to secondary sources, and such citations were removed during pre-marking processing. The apparent coverage drops even though the underlying research may have engaged with relevant authority. This is a methodological observation, not a vindication of Deep Research models' substantive citation practice. Even if hyperlinks were retained, they would still point primarily to secondary sources rather than to the primary authorities required by legal writing convention. Indeed, citing a greater number of unreliable sources should count against a model rather than in its favour. The format problem is therefore a symptom of a deeper source-selection problem, not an artefact of the marking protocol alone.

Taken together, the hallucination and accuracy–coverage findings indicate that the assessment of citation quality cannot be reduced to a single metric. A low hallucination count may reflect a well-disciplined model or an under-citing one, a high coverage figure may reflect genuine engagement with authority or a high tolerance for fabrication. Evaluating GenAI's suitability for legal tasks therefore requires the kind of disaggregated, multidimensional assessment adopted here, and any single-metric claim about model 'accuracy' in legal domains should be treated with scepticism.

3.5.2 The notable scholar pattern: 'Barbara McDonald'

A particularly illuminating pattern of hallucination was observed in relation to Professor Emeritus Barbara McDonald, whose actual body of work on Australian negligence law appears to have functioned as an 'attractor' for fabricated citations across multiple models and papers. In Paper 2, O3 (P2) generated the citation '*Barbara McDonald, "Re-thinking Causation in Australian Negligence Law" (2023) 46 UNSWLJ 1, 25–26.*' In Paper 3, GPT-4o (P3) generated '*Barbara McDonald, "Causation in Negligence" (2013) 35 Sydney Law Review 463.*' Neither publication exists.

A third instance was more subtle. In Paper 1, the model cited a genuine McDonald article: '*McDonald, "The Impact of the Civil Liability Legislation on Fundamental Policies and Principles of the Common Law" (2006) 14 Torts Law Journal 268 at 285*,' attributing to it the propositions that section 5D 'largely reflects the common law' while providing 'statutory guidance' rather than substantive change. The cited article is real; however, the quoted language could not be located on any page of it. Further, page 285 that is pinpointed by the model does not contain a clear relevance to the outlined discussion.

This form of hallucination, in which a real source is cited in support of propositions it does not contain, is arguably more insidious than outright invention of a source. The cited authority exists and only close inspection of the underlying text reveals that the quoted proposition is ungrounded. A reader who relies on the surface signal of a real citation, or is not familiar with the subject matter, is less likely to detect the fabrication.

Before generating the AI papers, we ran a pilot test on several models to minimise unexpected issues when generating all the AI answers. The pilot run in the Tort Law exam for O3 (P2) produced a further Barbara McDonald hallucination and its chain-of-thought trace is instructive. After searching the internet for the phrase 'Barbara McDonald causation section 5D 2023 ALJ,' the model proceeded immediately to state, 'I'm planning to reference key cases from [12]-[18] and cite relevant academic commentary, like *B McDonald, "Causation in Australian Negligence Law"* (2021).' The fabrication appears in the model's planning step, immediately after a search that would presumably not have returned a 2021 publication of that title. This sequence is inconsistent with the nature of internet search. The model appears to 'expect' that a McDonald article on causation and section 5D should exist, and where retrieval fails to return one, it provides a plausible-sounding substitute. The hallucination is therefore not random but systematic: the model fills a perceived lacuna in the literature with invented content that conforms to the expected pattern of such a source.

Grok 3 (P5) also made a reference to McDonald's publication: 'Academic commentary, such as McDonald (2006), argues section 5D's structure imposes a stricter "but for" application than common law's flexibility (e.g., material contribution in exceptional cases).' We could locate two 2006 publications by McDonald. In one of the two publications, there is a reference to the 'but for' test. However, the discussions do not support the claim made by Grok about stricter application of this test.

Table 10. Tort Law — case law hallucination divided by number of citations (repeated citations and hallucinations are counted). Lower is better

	Number of hallucinations/number of citations								
Question	**Claude Opus 4**	**O3**	**GPT-4o**	**Gemini (via AI Studio)**	**Grok 3**	**Perplexity Reasoning**	**Perplexity (Deep Research)**	**GPT-4.5 (Deep Research)**	**Grok 3 (Deeper Research)**
Q1	0/25	3/24	3/13	0/8	0/7	0/4	0/3	0/9	1/8
Q2	2/16	0/11	0/3	0/5	0/0	0/8	0/2	0/0	1/2
Q3	0/14	1/14	0/8	0/1	0/5	0/6	0/3	0/1	0/4
Q4	0/7	0/14	0/4	0/4	0/5	0/9	0/7	0/2	1/13
Total	**2/62 (3.23%)**	**4/63 (6.35%)**	**3/28 (10.71%)**	**0/18 (0%)**	**0/17 (0%)**	**0/27 (0%)**	**0/15 (0%)**	**0/12 (0%)**	**3/27 (11.11%)**

4. Discussion

The findings of this study suggest that law schools can no longer treat GenAI as a peripheral academic integrity issue.[39] The performance of current models, particularly those with Deep Research or reasoning capabilities shows that GenAI can now produce law exam responses that are not merely passable, but competitive with a substantial proportion of student work.

In Criminal Law, the models achieved an average score above the student cohort average and, in some cases, outperformed more than 90% of students. In Tort Law, performance was more uneven, but still sufficient for several models to obtain a pass,

[39] See also Hargreaves, "'Words Are Flowing Out Like Endless Rain Into a Paper Cup.'"

credit, distinction, or high distinction. This creates an immediate problem for non-invigilated assessments in which AI use is prohibited. A student who chooses to use GenAI may obtain a respectable grade, while students who comply with the rules may be placed at a comparative disadvantage.

This is not only an academic integrity problem. It is also an equity and employability problem. Grades remain relevant to clerkships, internships, graduate roles, scholarships, and other forms of professional opportunity. If non-invigilated assessments are retained in a context where GenAI is both accessible and capable of producing competent legal work, then students with stronger ethical commitments may be penalised for refusing to cheat. Conversely, students who outsource significant parts of their work to AI may be rewarded with marks that suggest a level of legal knowledge and reasoning they have not personally developed. Such an assessment system risks measuring access, willingness to use prohibited tools, and skill in avoiding detection, rather than legal understanding.

The results also challenge a more comforting assumption that GenAI may be useful for surface-level drafting but remains weak at legal reasoning. That assumption is becoming increasingly difficult to maintain. Critical analysis, long assumed to be a distinctively human strength and a clear AI weakness in the 2023 study, is improving, which is as impressive as it is concerning.[40] The models performed especially strongly on essay-style questions, likely because such questions allow them to draw on broad bodies of publicly available commentary, doctrine, and explanatory material. However, the more significant finding is that the models also performed well on problem-based hypothetical scenarios. This matters because legal education centres on the ability to appraise specialised skills, such as identifying legally relevant facts, selecting applicable principles, reasoning by analogy, and applying doctrine to complex circumstances to coherently and convincingly advise hypothetical parties. If GenAI can now produce competent responses to these tasks, the question for law schools is not whether students can use AI to complete assessments. The question is how legal education should respond. If a model can perform as well as a competent student, what is the distinctive contribution of the student, and how should we assess it?

One response would be to make all assessments invigilated and prohibit AI entirely.[41] This would protect the integrity of individual performance, but it would not prepare students for a profession in which AI tools are likely to be embedded in legal research, drafting, review, and advisory work.[42] Another response would be to allow AI freely in all assessments. That would reflect technological reality, but it risks allowing students to bypass the development of foundational legal knowledge and analytical judgment. Neither approach is sufficient on its own. Students need to learn law, and they need to learn how to work with AI. The assessment system must therefore incorporate both responses.[43]

The central educational objective should be what we describe as the Collaboration Surplus Test, i.e., student–AI collaboration should be assessed by whether, and to what extent, the human–AI output is superior to what either the student or the AI system could produce alone. Superiority may take the form of higher-quality work or delivering the same quality work faster or at lower cost. If the student's contribution does not improve the AI-generated output, the student's role is reduced to that of a passive user. In that scenario, the student remains vulnerable to professional displacement because the work product could be produced by the tool itself. Conversely, if the student's collaboration with GenAI produces no improvement over what the student could plausibly have produced independently, this may suggest that the student has not yet learned to use GenAI effectively, efficiently, or critically.

The Collaboration Surplus Test also cautions against relying too heavily on human attributes such as empathy, communication, and discretion. These qualities may lose some of their practical significance if AI systems can imitate them so convincingly that their performance becomes indistinguishable from human interaction. AI systems do not read, write, compare, or reason in the human sense; they process numerical representations. Yet they increasingly simulate these activities in ways that appear recognisably human. Some clients may be indifferent to whether their legal advice comes from a person or from a system that convincingly replicates human interaction, legal knowledge, and emotional responsiveness, especially if that service is available at a lower cost. Human attributes may remain professionally and ethically significant, but they are not necessarily a precondition for access to justice. Cost, by contrast, remains one of its most consequential determinants.[44]

As such, students need two distinct but connected forms of capability. First, students need to continue to develop foundational legal skills: doctrinal knowledge, legal research, case reading, statutory interpretation, issue identification, structured reasoning,

[40] Alimardani, "Generative Artificial Intelligence vs. Law Students."
[41] Nietzel, "UC Berkeley Law School Adopts New, Strict Ban On AI Use By Students."
[42] Zhang, Technology, Legal Education and Legal Profession in China and Australia.
[43] Alimardani, "Generative Artificial Intelligence vs. Law Students."
[44] Coumarelos, Legal Australia-Wide Survey.

and critical analysis. Without these skills, students cannot reliably distinguish when AI response is wrong, incomplete, misleading, or merely plausible. Second, it is imperative that students develop AI-collaboration skills: prompt design, context engineering, source verification, awareness of hallucination, understanding of model limitations, and the ability to avoid over-reliance.[45] AI competence does not substitute for legal skills. Rather, legal skill now depends on AI competence.

This point is critical considering the risk of verification drift. Alimardani's 2024 empirical study demonstrated that even when users understand that GenAI systems can hallucinate, they may gradually become less rigorous in verifying outputs, particularly where those outputs are legally fluent and appear credible.[46] In legal practice, this is especially dangerous because errors may be concealed behind authoritative language, misleading citations, or superficially coherent reasoning. Empirical work on the misuse of GenAI across jurisdictions illustrates that this risk is not merely hypothetical. In one reported instance, a lawyer who had previously been found to have submitted hallucinated material to a court went on to make further hallucinated submissions only months later in another case.[47] The present study similarly found that some models produced strong legal analysis while also generating hallucinated or unsupported citations. This combination is concerning. A weak answer is relatively easy to detect; a strong-looking answer containing subtle citation or reasoning errors is more dangerous because it may pass superficial review.

Context engineering may reduce some of these risks, but it will not remove them. The models in this study were not given subject-specific teaching materials, and it is likely that their performance would improve if they were connected to curated legal databases, lecture materials, textbooks, or case repositories. Retrieval-augmented generation (RAG) may assist by grounding outputs in large datasets.[48] However, RAG systems also have limitations. They may retrieve irrelevant materials, miss important sources, treat weak secondary sources as equivalent to primary law, or fail to integrate retrieved material properly into legal reasoning.[49] The issue, therefore, is not simply whether a model has access to more information. It is whether students can evaluate the quality, relevance, authority, and use of that information.

The unevenness of model performance also cautions against generalisations about AI competence. The models did not perform uniformly across subjects, question types, or assessment criteria. Some models were strong in Criminal Law but weak in Tort Law. Some performed well on critical analysis but poorly on citation. Some Deep Research models reduced hallucination but also produced thinner coverage of relevant authorities. This jagged performance means that students cannot safely assume that a model which performs well in one task will perform well in another.[50] More concerningly, model performance may also be temporally jagged. Capabilities may improve in some respects while deteriorating in others across model updates.[51] Legal education must therefore prepare students for unstable and uneven AI performance, not a linear trajectory of continuous improvement.

Drawing on these findings together with the educational literature discussed above, we recommend an assessment architecture that combines three forms of evaluation, rather than relying on any single assessment mode. The first category is invigilated assessment in which AI use is prohibited and precluded. Its purpose is to verify that students have developed foundational legal knowledge and independent analytical capacity without AI support.[52] This may include traditional exams, in-class problem questions, oral assessments, timed drafting exercises, or supervised legal analysis tasks. Such assessments are necessary because students cannot supervise AI effectively if they lack the underlying knowledge needed to detect errors.[53] Invigilated assessment is therefore not a nostalgic return to pre-AI education. It is a necessary foundation for responsible AI use. We recommend that during the first year of law school, AI-free invigilated tasks should form the bulk of assessments. The rationale for this recommendation is to prioritise the development of foundational legal skills before students are assessed on AI-supported performance, while reducing the risk of over-reliance on GenAI.

[45] Head, "Assessing Law Students in a GenAI World to Create Knowledgeable Future Lawyers"; Guo, "A Transformative AI-Augmented Lawyering Skills Approach for the Vis Moot."
[46] Alimardani, "Borderline Disaster."
[47] Legg, "The Promise and the Peril of the Use of Generative Artificial Intelligence in Litigation."
[48] Gao, "Retrieval-Augmented Generation for Large Language Models."
[49] Chen, "Benchmarking Large Language Models in Retrieval-Augmented Generation."
[50] Dell'Acqua, "Navigating the Jagged Technological Frontier."
[51] Alimardani, "GenAI and the Mirage of Personalised Learning for All"; Lee, "The Impact of Generative AI on Higher Education Learning and Teaching."
[52] Head, "Assessing Law Students in a GenAI World to Create Knowledgeable Future Lawyers."
[53] Abbasi, "Responsible Legal Augmentation."

The second category is AI-collaborative assessment. In this model, AI use is permitted and should be required, not optional.[54] Assessment should move away from focusing solely on the final product and instead evaluate both the process and the final product.[55] The assessment should therefore not mark only an essay, advice, memorandum, or report. It should also assess the student's process of collaboration: what they asked the model to do, what sources they used, how they verified the output, which suggestions they accepted or rejected, how they responded to errors or hallucinations, and how their own legal reasoning improved the final submission.[56] This approach recognises that, in professional practice, value increasingly lies not merely in producing an output, but in supervising, directing, and improving AI-generated work. It shifts the assessment focus from concealment to demonstration. Students are not rewarded for pretending that AI was not used. They are rewarded for showing that they understood where they should use AI, and whether they used it critically, ethically, and effectively.

The third category is the Relay Model, a staged human–AI handoff which combines AI and AI-free stages within a single assessment. In an AI-first relay, students first use AI to generate a draft or research output, and then, in an invigilated setting without AI access, they must critique, verify, correct, and improve that output. This directly tests whether students can identify weak reasoning, hallucinations, unsupported claims, poor authorities, and missed issues. In a human-first relay, students first produce their own work under invigilated conditions and then use AI to review and improve their drafts. This tests whether students can use AI to elevate rather than replace their own independent legal analysis. The Relay Model is valuable because it resembles plausible professional uses of GenAI, where lawyers may both review AI-generated drafts and use AI to refine their own work. In practice, this process may be iterative rather than linear. A lawyer may use AI to produce an initial draft, review and revise it, ask the AI for further analysis or alternative perspectives, and then review and revise the output again. The value lies not in a single interaction with AI, but in the lawyer's ability to supervise and improve the work through repeated cycles of human-AI collaboration.

Our broader recommendation is that legal education assessment should be either supervised where AI is not allowed or deliberately designed around AI collaboration where AI is allowed. Ambiguous middle-ground assessments are increasingly difficult to justify. If an assessment is non-invigilated and AI is prohibited, enforcement will often be unreliable and may reward dishonest behaviour. If AI is permitted but the assessment focuses only on the final product and does not evaluate how students collaborated with AI, students may use AI as a substitute for learning rather than as a tool to enhance their own reasoning. AI-enabled assessments should therefore assess both the quality of the final work and the student's collaboration process, including how they prompted the system, verified outputs, identified errors, exercised legal judgment, and improved the AI-generated material. The more coherent approach is to align the assessment mode with the learning objective. Where the objective is independent legal reasoning, the assessment should be invigilated. Where the objective is professional AI capability, the assessment should require and assess AI collaboration.[57]

Contrary to the promise that AI would make intellectual work easier, it is likely to make it more difficult for students in some respects. To remain relevant, students will need to understand the law well enough to outperform or improve AI outputs, and they will need to understand AI well enough to use it without becoming dependent on it. The benchmark is no longer simply whether a student can produce a competent answer. It is whether the student can produce, verify, and improve legal work in an environment where AI can also produce competent answers. That is a higher standard, but it is the standard that legal education must now confront.

Future research should examine how models perform when connected to legal databases and professional legal AI platforms. It is plausible that such systems will outperform general-purpose models, particularly in citation accuracy and source retrieval. However, their performance should not be assumed. Legal databases, RAG pipelines, and specialised tools require empirical evaluation, especially across different legal subjects, jurisdictions, and assessment types. Future studies should also examine whether newer models become more accurate overall while still exhibiting jagged or temporally unstable performance in particular legal capabilities. The benchmark figures should therefore be read as a time-specific snapshot of the capabilities of the models tested in 2025. The study's more durable contribution lies in the assessment-design argument those results motivate, rather than in the continued currency of any particular model or percentile result.

[54] One downside of mandatory AI use is that some students may regard the use of AI as unethical, because the works of scholars, artists, and other creators may have been used to develop AI models without their consent. There are other concerns such as the impact of AI on the environment and the cost of resources such as electricity. Some studies have also raised concerns about students' anxiety surrounding the use of GenAI. See Oberg, "Feeling AI."

[55] Alimardani, "Borderline Disaster."

[56] Winstone, "Black Box Assessment"; Guo, "A Transformative AI-Augmented Lawyering Skills Approach for the Vis Moot"; Lodge, Assessment Reform for the Age of Artificial Intelligence.

[57] Alimardani, "Generative Artificial Intelligence vs. Law Students."

5. Conclusion

This study set out to re-examine how GenAI performs in university law assessments two years after the 2023 baseline. The results show that the tested GenAI configurations produced exam responses that were competitive with student work. Across Criminal Law and Tort Law, contemporary models no longer merely pass. They compete with and frequently exceed a substantial proportion of the student cohort. Despite the jagged and unstable nature of AI models' capability across subjects, question types, and successive updates, this study suggests that the gap between AI and law students is closing rapidly and that critical legal analysis can no longer be treated as a dependable point of human distinction. The core conclusion is not that law schools should resist AI, nor that they should embrace it without constraint. The conclusion is that assessment must be redesigned around the reality that GenAI can now perform many law assessment tasks at a competent or better level. Law schools must protect the development of foundational legal skills while also teaching students to collaborate with AI in ways that produce genuine human value-add (Collaboration Surplus Test). The aim must be to produce graduates who can collaborate with GenAI to produce work superior to what either the student or the AI could produce independently. To achieve this, assessments must use a combination of the following formats. Invigilated and AI-free assessments verify independent legal knowledge and skills. AI-collaborative assessments evaluate process as well as product, teaching students to use AI to their advantage and not as a tool to outsource their understanding. Relay assessments connect these capabilities through staged tasks: students either produce an initial response independently under invigilated conditions and then use AI to improve it, or critique and revise an AI-generated response in an invigilated, AI-free setting. Together, these formats assess both what students can do independently and the value they contribute when working with AI.

AI Disclosure

After the manuscript was finalised, AI tools (ChatGPT-5.5 and Claude Opus 4.8) were used to identify errors, including conflicting claims, table inconsistencies, and typographical issues. The GenAI models identified a small number of grammatical and expression-related issues, as well as one table error. Each suggestion was carefully reviewed before being applied.

Bibliography

Abbasi, Zubair. "Responsible Legal Augmentation: Integrating Generative AI into Legal Practice." *Law, Technology and Humans* 7, no 3 (2025): 5–13. https://doi.org/10.5204/lthj.4200.

Abhyankar, Reyna, Qi Qi and Yiying Zhang. "Osworld-Human: Benchmarking the Efficiency of Computer-Use Agents." Preprint, arXiv, June 19, 2025. https://arxiv.org/abs/2506.16042.

Alimardani, Armin. "Borderline Disaster: An Empirical Study on Student Usage of GenAI in a Law Assignment." *IEEE Transactions on Technology and Society* 6, no 2 (2025): 210–19. https://doi.org/10.1109/TTS.2025.3540978.

Alimardani, Armin. "Generative Artificial Intelligence vs. Law Students: An Empirical Study on Criminal Law Exam Performance." *Law, Innovation and Technology* 16, no 2 (2024): 777–819. https://doi.org/10.1080/17579961.2024.2392932.

Alimardani, Armin and Emma A. Jane. "GenAI and the Mirage of Personalised Learning for All." *Law, Technology and Humans* 7, no 2 (2025): 63–88. https://doi.org/10.5204/lthj.3764.

Bell, Felicity and Justine Rogers. "Becoming, Doing, Being: GenAI and the Promise of Professional Identity in Law." *Law, Technology and Humans* 7, no 3 (2025): 62–93. https://doi.org/10.5204/lthj.4120.

Borges, Beatriz, Negar Foroutan, Deniz Bayazit et al. "Could ChatGPT Get an Engineering Degree? Evaluating Higher Education Vulnerability to AI Assistants." *Proceedings of the National Academy of Sciences* 121, no 49 (2024): e2414955121. https://doi.org/10.1073/pnas.2414955121.

Burgess, Paul, Iwan Williams, Lizhen Qu and Weiqing Wang. "Using Generative AI to Identify Arguments in Judges' Reasons: Accuracy and Benefits for Students." *Law, Technology and Humans* 6, no 3 (2024): 5–22. https://doi.org/10.5204/lthj.3637.

Chen, Jiawei, Hongyu Lin, Xianpei Han and Le Sun. "Benchmarking Large Language Models in Retrieval-Augmented Generation." *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no 16 (2024): 17754–62. https://doi.org/10.1609/aaai.v38i16.29728.

Choi, Jonathan H. and Daniel Schwarcz. "AI Assistance in Legal Analysis: An Empirical Study." *Journal of Legal Education* 73, no 2 (2025): 384–420. https://dx.doi.org/10.2139/ssrn.4539836.

Coumarelos, Christine, Deborah Macourt, Julie People et al. *Legal Australia-Wide Survey: Legal Need in Australia*. Law and Justice Foundation of New South Wales, 2012.

Dahl, Matthew, Varun Magesh, Mirac Suzgun and Daniel E. Ho. "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models." *Journal of Legal Analysis* 16, no 1 (2024): 64–93. https://doi.org/10.1093/jla/laae003.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality." *Organization Science* 37, no. 2 (2026): 403–23. https://doi.org/10.1287/orsc.2025.21838.

Guo, Peng and Alex Wan. "A Transformative AI-Augmented Lawyering Skills Approach for the Vis Moot." *Vindobona Journal* 27, no 2 (2023): 127–34.

Gao, Yunfan, Yun Xiong, Xinyu Gao et al. "Retrieval-Augmented Generation for Large Language Models: A Survey." Preprint, arXiv, December 18, 2023. https://arxiv.org/abs/2312.10997.

Hargreaves, Stuart. "ChatGPT and Other Generative AI Tools as University Teaching Aids." *Legal Education Review* 35, no 1 (2025): 55–74. https://doi.org/10.53300/001c.137110.

Hargreaves, Stuart. "'Words Are Flowing Out Like Endless Rain Into a Paper Cup': ChatGPT & Law School Assessments." *Legal Education Review* 33, no 1 (2023): 69–105. https://doi.org/10.53300/001c.83297.

Head, Amanda and Sonya Willis. "Assessing Law Students in a GenAI World to Create Knowledgeable Future Lawyers." *International Journal of the Legal Profession* 31, no 3 (2024): 293–310. https://doi.org/10.1080/09695958.2024.2379785.

Kommers, Cody, Eamon Duede, Julia Gordon et al. "Why Slop Matters." *ACM AI Letters* 1, no 1 (2026): 1–6. https://doi.org/10.1145/3786777.

Landy, Nicole. "Early PLT Student Perceptions of the Integration of Generative Artificial Intelligence in Legal Education." *Law, Technology and Humans* 7, no 3 (2025): 33–51. https://doi.org/10.5204/lthj.4031.

Lee, Daniel, Matthew Arnold, Amit Srivastava et al. "The Impact of Generative AI on Higher Education Learning and Teaching: A Study of Educators' Perspectives." *Computers and Education: Artificial Intelligence* 6 (2024): 100221. https://doi.org/10.1016/j.caeai.2024.100221.

Legg, Michael, Vicki McNamara, and Armin Alimardani. "The Promise and the Peril of the Use of Generative Artificial Intelligence in Litigation." *University of New South Wales Law Journal* 48, no. 4 (2025): 1196–1235. https://doi.org/10.53637/cjxx1348.

Lindkvist, Annica, Sophie Curbo and Catharina Hultgren. "Choosing Not to Use Generative AI in Higher Education: A Mixed Methods Study of Student Reasoning, Assessment Design, and AI Literacy." *Frontiers in Education* 11 (2026): 1813306. https://doi.org/10.3389/feduc.2026.1813306.

Lodge, Jason, Sarah Howard, Margaret Lynn Bearman and Phillip Dawson. *Assessment Reform for the Age of Artificial Intelligence*. Tertiary Education Quality and Standards Agency, 2023.
Lodge, Jason M., Kate Thompson and Linda Corrin. "Mapping Out a Research Agenda for Generative Artificial Intelligence in Tertiary Education." *Australasian Journal of Educational Technology* 39, no 1 (2023): 1–8. https://doi.org/10.14742/ajet.8695.
Madden, Raul. "The Dignity of Legal Education vs Artificial Intelligence in Australian Law Schools." *ANU Journal of Law and Technology* 6, no 1 (2025): 7–50.
Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning and Daniel E. Ho. "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools." *Journal of Empirical Legal Studies* 22, no 2 (2025): 216–42. https://doi.org/10.1111/jels.12413.
Marr, Bernard. "6 Skills You Need to Stand Out in the Age of AI." Fast Company, March 28, 2023. https://www.fastcompany.com/90871901/6-skills-you-need-to-stand-out-in-the-age-of-ai.
Mawn, Mary V. and Jacqueline M. Washington. "Using Artificial Intelligence as a Mindtool to Foster Critical Thinking in Biology Education." *Journal of Microbiology & Biology Education*, ahead of print, May 11, 2026: e00048-26. https://doi.org/10.1128/jmbe.00048-26.
Nietzel, Michael T. "UC Berkeley Law School Adopts New, Strict Ban On AI Use By Students." Forbes, May 22, 2026. https://www.forbes.com/sites/michaeltnietzel/2026/05/22/uc-berkeley-law-school-adopts-new-strict-ban-on-ai-use-by-students/.
Oberg, Glenys, Yifei Liang, Margaret Bearman, Tim Fawns, Michael Henderson and Kelly E. Matthews. "Feeling AI: Circulating Emotions, Institutional Climates, and Moral Boundaries in Student Use of AI." *Higher Education*, ahead of print, March 28, 2026. https://doi.org/10.1007/s10734-026-01658-6.
OpenAI, Achiam, Josh, Steven Adler, Sandhini Agarwal et al. "GPT-4 Technical Report." Preprint, arXiv, March 15, 2023. https://arxiv.org/abs/2303.08774.
Open AI Jaech, Aaron, Adam Kalai, Adam Lerer et al. "OpenAI O1 System Card," Preprint, arXiv, April 30, 2026, https://doi.org/10.48550/arXiv.2412.16720.
Steinacker-Olsztyn, Nicolas, Devashish Gosain and Ha Dao. "Is Misinformation More Open? A Study of Robots.Txt Gatekeeping on the Web." In *WWW '26: Proceedings of the ACM Web Conference 2026*, 7507–16. Association for Computing Machinery, 2026. https://doi.org/10.1145/3774904.3792625.
Sun, Jingbo, Wenyue Chong, Songjun Tu et al. "AutoSearch: Adaptive Search Depth for Efficient Agentic RAG via Reinforcement Learning." In *Findings of the Association for Computational Linguistics: ACL 2026*, edited by Maria Liakata, Viviane P. Moreira, Jiajun Zhang and David Jurgens, 28059–79. Association for Computational Linguistics, 2026. https://doi.org/10.18653/v1/2026.findings-acl.1399.
The Law Society of New South Wales. *Solicitor Capability Framework*. 2025. https://www.lawsociety.com.au/sites/default/files/2025-08/LS4645_CPD_CapabilityFramework_2025-08-29.pdf.
Thompson, Kate, Linda Corrin and Jason M. Lodge. "AI in Tertiary Education: Progress on Research and Practice." *Australasian Journal of Educational Technology* 39, no 5 (2023): 1–7. https://doi.org/10.14742/ajet.9251.
Winstone, Naomi E., Karen Gravett and Sam Elkington. "Black Box Assessment: Rethinking Integrity and Learning for a Time of Generative AI." *Assessment & Evaluation in Higher Education*, ahead of print, April 28, 2026. https://doi.org/10.1080/02602938.2026.2661365.
Yuvaraj, Joshua. "The Verification-Value Paradox: A Normative Critique of Gen AI in Legal Practice." *Monash University Law Review,* ahead of print, 2026.
Zhang, Shu, Jie Luo and Peng Guo. *Technology, Legal Education and Legal Profession in China and Australia: Opportunities and Challenges*. International and Comparative Law in the Asia Pacific. Springer Nature Singapore, 2024. https://doi.org/10.1007/978-981-96-1639-8.

Cases

Rogers v Whitaker (1992) 175 CLR 479.