GenAI and the Mirage of Personalised Learning for All

Armin Alimardani
University of Wollongong, Australia
Emma A. Jane
UNSW Sydney, Australia

Abstract

Generative artificial intelligence (GenAI) is gaining attention in education, particularly for its potential to personalise learning. However, studies have yet to assess its effectiveness and identify its limitations within controlled educational contexts across various disciplines and GenAI models. This study aims to help fill this gap by evaluating 'SmartTest', a GenAI chatbot designed by the authors to prompt questions, offer immediate feedback, and stimulate critical thinking through conversational nudges. Over five test cycles, SmartTest was used in a criminal law course at an Australian university. The results revealed SmartTest struggled with complex structured problem-solving exercises. While SmartTest showed some promise in aiding short-answer question learning, it remained limited and prone to inaccuracies. These findings highlight concerns that inflated expectations of GenAI could mislead educators and investors about its capabilities to deliver economically viable and quality personalised learning.

***Keywords**:* Generative artificial intelligence; law; higher education; chatbot; ChatGPT; GPT-4.

1. Introduction

The rapid rise of generative artificial intelligence (GenAI) tools that mimic human conversation is posing significant challenges for educators. For instance, when OpenAI's ChatGPT (GPT-3.5) was launched as a prototype in late 2022, it performed in the bottom 10% of United States bar exam participants. Just four months later, OpenAI claimed that GPT-4, an improved version of the model, scored higher than 90% of human test takers.[1] Such GenAI capabilities have prompted some educators to advocate for banning artificial intelligence (AI) in classrooms and returning to traditional pen-and-paper exams to combat academic dishonesty.[2] Others argue it is unrealistic to expect students to avoid GenAI tools, given their widespread availability and relevance in professional settings. They propose integrating GenAI into assessment tasks.[3] Meanwhile, a third group sees GenAI as an opportunity to democratise education and enhance personalised learning.[4]

---

[1] OpenAI, "GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses." Some studies questioned GPT-4's capability to perform better than 90% of bar exam test takers. See Alimardani, "Generative Artificial Intelligence vs. Law Students."
[2] Cassidy, "Australian Universities to Return to 'Pen and Paper' Exams after Students Caught Using AI to Write Essays."
[3] Foung, "Reinventing Assessments with ChatGPT and Other Online Tools"; Furze, "The AI Assessment Scale (AIAS) in Action"; Alimardani, "Borderline Disaster."
[4] Yan, "Generative Artificial Intelligence and Human Learning"; Calo, "Towards Educator-Driven Tutor Authoring"; Grover, "Next-Generation Education"; Guettala, "Generative Artificial Intelligence in Education"; Guo, "Enhancing Constructivist Learning"; Nikolic, "Prompt Potential"; Burgess, "Using Generative AI to Identify Arguments in Judges' Reasons."

While the use of GenAI as an intelligent support chatbot in classrooms is widely discussed,[5] the accuracy and reliability of such tools have yet to be tested in controlled educational settings in various domains and against different GenAI models.[6] Further, despite the increasing accessibility of GenAI to developers for AI-powered applications, the development of this transformative technology remains predominantly in the hands of science, technology, engineering, and mathematics (STEM) disciplines.[7] This leaves many scholars in the humanities and social sciences (HASS) engaging with AI only in abstract terms. The proprietary nature of corporate-developed chatbots further widens this gap, raising concerns about transparency, data privacy, and alignment with pedagogical values.

This study aims to empirically evaluate the capabilities of GenAI through 'SmartTest', a chatbot developed by the authors (both of whom are situated in HASS disciplines). SmartTest is a Socrative chatbot which offers immediate feedback and uses conversational nudges to encourage critical thinking.[8] It was piloted across five test cycles in a criminal law course at the University of Wollongong (UOW) in late 2023. Findings reveal that GenAI struggles with complex, multi-step problem-solving tasks but shows more potential in handling simpler, short-answer questions, providing useful hints when students err. However, some limitations remain, including the potential for students to exploit the chatbot for unintended uses, such as completing unrelated class activities. A student survey indicated appreciation for SmartTest's ability to reduce anxiety and guide students towards correct answers. Despite this, around half of students still preferred feedback from human instructors, even if delayed by over a day.

This study offers three primary findings. First, there is a productivity paradox regarding the use of GenAI chatbots within small- to medium-sized cohorts. Drafting exercises for GenAI chatbots is time-consuming and requires rigorous testing to identify and address inaccurate AI outputs. This challenges the notion that GenAI provides a shortcut to personalised education, as it may ultimately increase rather than reduce the workload for educators. Second, GenAI models demonstrate inconsistency. Even under identical conditions, these models may generate accurate outputs in one interaction and erroneous outputs in another. Third, given that the current GenAI models typically require human oversight to verify their accuracy,[9] we would discourage the use of such chatbots as a trustworthy personalised tutor where definitive answers or precise solutions are required. However, for questions without a single correct answer or where multiple interpretations are possible, GenAI chatbots could be beneficial in supporting students' critical thinking skills.

This study's framework focuses on an exploratory evaluation of GenAI as an educational chatbot. Rather than concentrating extensively on a single aspect of GenAI's implementation in educational settings with detailed methodological and statistical analyses, this study examines multiple dimensions of the topic. Its primary goal is to identify emerging themes and point out key areas for future research.

Following this introduction, the article consists of five main sections and a conclusion. Section 2 outlines the methodology, focusing on the development of SmartTest and the design of the test cycles. Section 3 presents the results of students' interactions with SmartTest, identifying both strengths and concerns. Building upon these findings, Section 4 discusses implications, such as the potential for student misuse of AI chatbots and the observation that on certain tasks, newer GenAI models integrated into SmartTest demonstrate poorer performance compared to earlier versions. Section 5 analyses student survey responses about their experiences with SmartTest, and Section 6 addresses the study's limitations.

2. Methodology

Exploring GenAI as a personal tutor requires systematically examining its effectiveness in practical educational settings. One approach could involve students using ChatGPT to generate questions on specific topics. However, concerns arise about ChatGPT's subject matter knowledge, its tendency to 'hallucinate'—producing nonsensical or inaccurate information[10]—and the risk of providing legal rules from irrelevant jurisdictions. It is also uncertain whether ChatGPT can guide students through

[5] Yan, "Generative Artificial Intelligence and Human Learning"; Calo, "Towards Educator-Driven Tutor Authoring"; Grover, "Next-Generation Education"; Guettala, "Generative Artificial Intelligence in Education"; Guo, "Enhancing Constructivist Learning"; Kerlyl, "Bringing Chatbots into Education."
[6] Some studies investigated the use of GenAI tutors in educational settings but did not clearly examine instances where these tools provided erroneous or incorrect feedback. See Vanzo, "GPT-4 as a Homework Tutor can Improve Student Engagement and Learning Outcomes"; Henkel, "Effective and Scalable Math Support"; Chen, "Intelligent Tutor"; Ma, "Socratic ChatGPT"; Binhammad, "Investigating how Generative AI can Create Personalized Learning Materials Tailored to Individual Student Needs."
[7] Calo, "Towards Educator-Driven Tutor Authoring."
[8] Dillon, Teaching Psychology and the Socratic Method.
[9] Passi, "Overreliance on AI Literature Review"; Achiam, "GPT-4 Technical Report."
[10] Achiam, "GPT-4 Technical Report," 46.

corrective steps when they answer incorrectly, rather than simply offering the correct answers. Additionally, ChatGPT may lack the ability to deliver a personalised learning experience that adapts to a student's knowledge level. An alternative strategy that was used in this study involves embedding educational instructions, questions, and answer guides drafted by educators into the GenAI prompt. This method improves content accuracy and consistency but requires students to input the prompt, exposing them to the answers, which could reduce cognitive engagement. Therefore, to address this issue, we used a platform that can implement and conceal the prompt within embedded material and provide a more effective learning environment.

2.1 Safe-to-fail AI Project: Designing SmartTest

The authors initiated the Safe-to-Fail AI[11] project in the second quarter of 2022 to promote and implement the democratisation of emerging technologies across academic disciplines. Our goal was to design and deploy AI-integrated educational and research tools that would be free and publicly accessible, allowing educators to explore them without needing technical expertise. This ambitious project was launched more than six months before ChatGPT's release, when GenAI was neither in the spotlight nor widely accessible. We used a third-party platform[12] to develop these tools[13] and officially launched the project in September 2022.[14]

For this study, we created SmartTest, an educational chatbot with features designed to align with educational objectives. The chatbot allows educators to pose pre-determined questions to students, ensuring that the topics and focus areas match learning goals. Feedback is generated based on educator-drafted answers to provide a coherent learning path. Additionally, the chatbot is programmed to engage with students conversationally, encouraging critical thinking and guided discovery rather than simply offering solutions. To keep the prompt—including questions, answers, and instructions—hidden from students, we implemented a 'system prompt' section. A system prompt is an internal set of instructions provided to an AI chatbot, guiding its behaviour, tone, and the types of responses it generates, which is usually hidden from users.[15] For instance, in a legal chatbot, the system prompt might include instructions to maintain a formal tone, focus on specific legal jurisdictions, and avoid irrelevant topics.

Image 1 illustrates an example of how a student might initiate a conversation with SmartTest, which is configured to follow pre-instructed sets of questions. The student cannot see the instructions and correct answer under the system prompt, ensuring that the learning process remains authentic and challenging. Image 2 shows how educators can access the system prompt section to outline their desired instructions, questions and answers. This feature allows educators to customise the chatbot's functionality and ensures it aligns with their educational objectives while maintaining the confidentiality of the prompts from students.

[11] Alimardani, "Safe-to-fail AI."
[12] Streamlit, "A Faster Way to Build and Share Data Apps."
[13] Through OpenAI API.
[14] Alimardani, "Safe-to-fail AI."
[15] OpenAI, "System Messages."

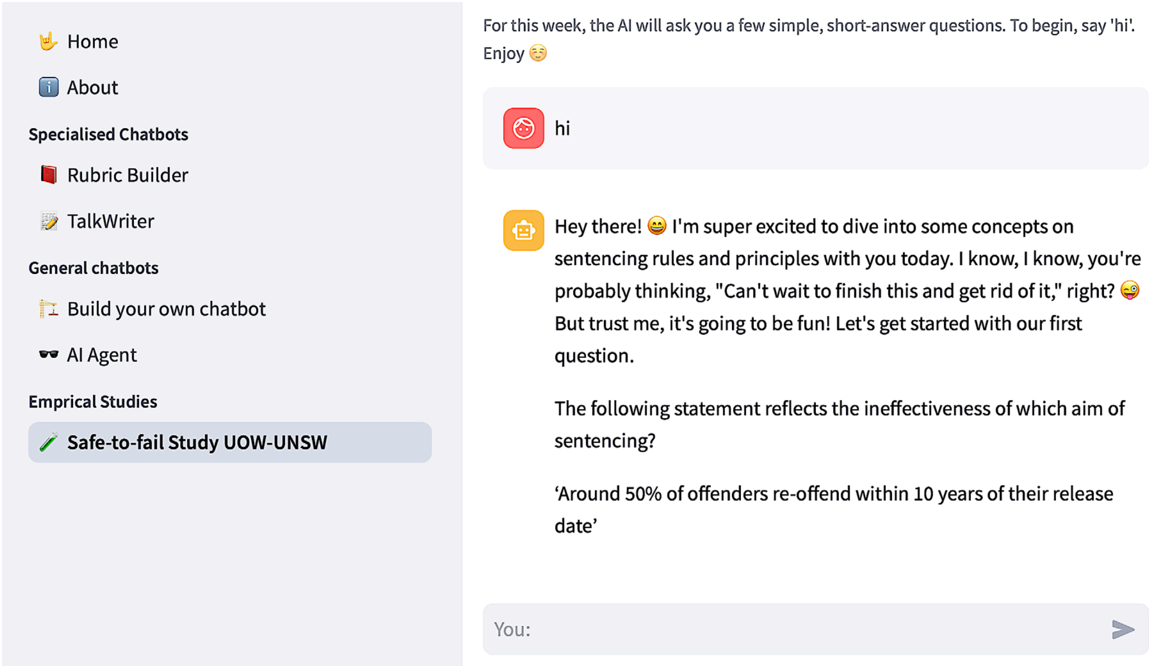

Image 1. Student interaction with SmartTest with hidden system prompt.
Credit: **Screenshot of the Streamlit platform interface. Reproduced with permission from Snowflake Inc.**[16]

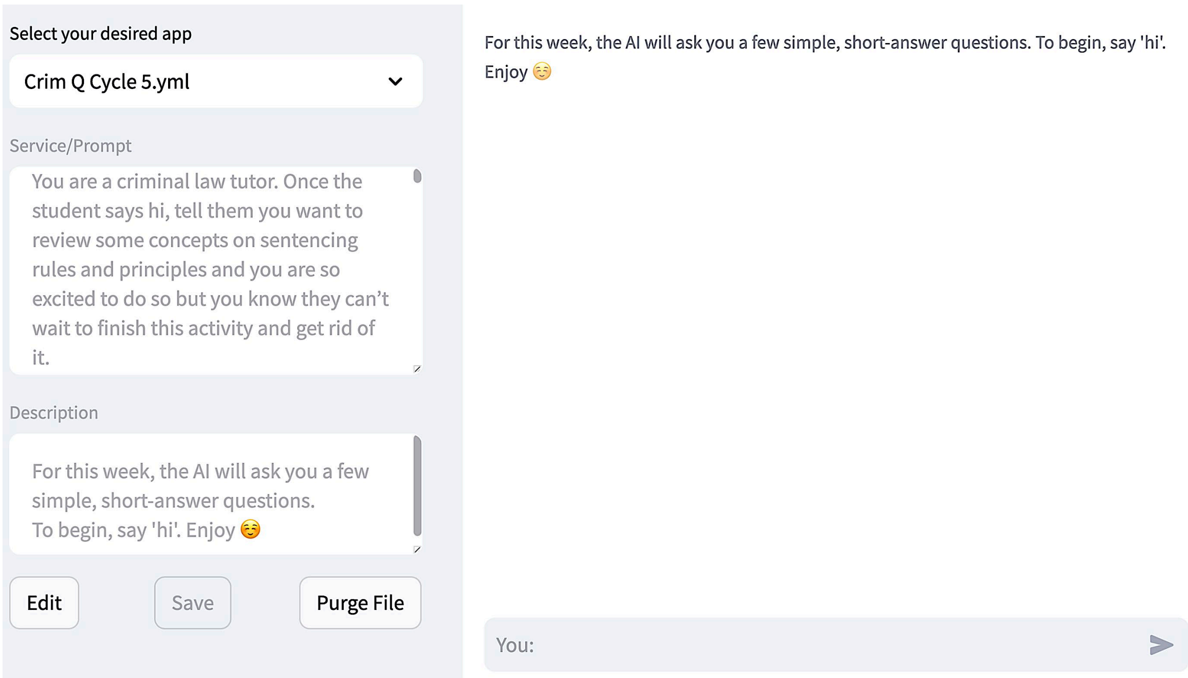

Image 2. Access to the system prompt for educators.
Credit: **Screenshot of the Streamlit platform interface. Reproduced with permission from Snowflake Inc.**[17]

[16] https://streamlit.io
[17] https://streamlit.io

2.2 Study Design

We conducted this study during Semester 2, 2023, in Criminal Law and Procedure B tutorials at the UOW over five test cycles in weeks 5, 7, 9, 10,[18] and 12 of the semester. In the design of the test cycles, we aimed to systematically explore varying levels of task complexity and question formats for GenAI. The design of our study, as well as the broader project, was informed by the principles of the 'safe-to-fail' approach. This framework encourages small-scale, exploratory experiments in complex socio-technical environments to minimise unforeseen risks.[19] In this context, considering that the use of GenAI as an educational chatbot is still in its early stages and its associated risks are not yet fully understood, we aimed to involve approximately 30 students in each test cycle. At the School of Law, UOW, core subject classes have a maximum enrolment capacity of 25 students. As we expected only partial participation from each class, we estimated that two to three classes would be sufficient to reach the target. However, we anticipated challenges, such as tutors potentially being unable to conduct the experiment due to time constraints. To account for such issues, we planned to run the test cycles in three to four classes.

Since student participation was voluntary and anonymous, it was not possible to determine the total number of participants in this study. Some students may have participated in certain test cycles but not in others. Consequently, we could only calculate the number of students participating in each test cycle separately. See Table 1 for details on participant numbers across each test cycle.

Table 1. Number of participants in each test cycle

Cycle number	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5
Number of participants/conversations*	39	43	38	26	32
Number of classes involved	3	4	4	2**	2**

* This table includes only interactions where students provided at least two inputs, such as continuing the conversation beyond an initial greeting (e.g., after saying 'hi').
** In both Cycles 4 and 5, three classes were organised to participate; however, in each cycle, one tutor was unable to conduct the experiment. Despite this, both cycles generated sufficient data to evaluate the capabilities of SmartTest for the purpose of this study.

To address the consequences of AI errors, such as hallucinations, and to ensure students were not given inaccurate information by SmartTest, tutors reviewed the exercises after students had completed them and demonstrated the correct answers to the class. All student interactions with SmartTest were recorded in a Google Docs document. The study adhered to UOW ethics committee guidelines: participation was voluntary and anonymous; SmartTest sessions were capped at 10 minutes to minimise potential disruption to the tutors' lesson plans; to avoid any perceived pressure, none of the authors were present in tutorials during the tests; and the committee restricted the authors from receiving direct feedback from the students or their tutors. To address the latter constraint, SmartTest was configured to recognise the end of the test and prompt students to provide feedback, allowing for anonymous and unbiased data collection in each test cycle. However, many students did not finish the test or ignored the feedback questions. Consequently, a feedback survey was conducted in Week 13 of the semester.

2.3 Test Cycle Format Summary

The system prompts in the test cycles included three main sections: (1) Instructions—SmartTest's role and the steps it should take when interacting with students. For example, if a student asks an unrelated question, SmartTest refuses to answer and redirects them to the test cycle question; (2) The Questions; and (3) The Answer Guide. We drafted all the SmartTest questions for these test cycles from the same week's teaching material to boost student engagement with SmartTest.

In the first three cycles, we used short problem scenarios—hypothetical criminal case questions. For instance, students were asked whether the prosecution could prove an offence based on the information provided in the hypothetical case. To answer problem scenario questions, students needed to identify the relevant elements of the offence and argue the strengths and weaknesses of the prosecution's case. This required listing the relevant legal authorities and explaining how facts in the hypothetical scenario might prove or disprove elements of the offence. To initially understand SmartTest's capabilities, we drafted the first cycle with a short hypothetical problem scenario that included simple instructions and an answer guide. We

[18] There was a one-week break between weeks 9 and 10.
[19] Snowden, "A Leader's Framework for Decision Making"; Garbett, "Safe-to-Fail Experiments."

introduced more complicated problem scenarios, instructions, and answer guides in the second and third test cycles. These cycles aimed to enable the student to lead the conversation and arrive at the correct answer step by step. If a student answered incorrectly, we instructed SmartTest to generate a hint, based on the question and the correct answer, rather than providing the correct answer outright.

In Cycles 4 and 5, we drafted several short-answer questions, with SmartTest taking the lead by asking the questions and students providing the answers. We introduced a hint section for each question, instructing SmartTest to use this information to guide students when they answered incorrectly.

In Cycle 5, we enhanced personalised learning by configuring SmartTest to adapt questions based on students' responses. For example, if a student answered a question incorrectly, SmartTest would pose another pre-drafted question on the same topic to ensure comprehension of the material. We also incorporated 'content' sections that provided general information about each question's topic. This prevented SmartTest from relying on its own prior knowledge when students asked follow-up questions or requested further elaboration. Instead, SmartTest used the content as a reliable source of information to address student inquiries.

2.4 AI Engines and Prompt Engineering

The GenAI model we initially used for SmartTest in 2022 was GPT-3 (DaVinci 2 Engine), the predecessor to ChatGPT (GPT-3.5). Similar to ChatGPT, this model had several weaknesses, including challenges in following instructions and reasoning effectively. In March 2023, OpenAI released GPT-4 to a small group of early adopters for testing a few months before making it available to the general public. We were fortunate to be granted early access[20] and implemented GPT-4 in the Safe-to-Fail AI platform. Given its enhanced capabilities compared to previous models, we tested it in classroom settings and prepared the platform for use before the second semester of 2023 at UOW.

OpenAI provides developers with two types of AI models: continuous and static.[21] Static models remain unchanged from their release date. These models allow for performance comparisons with other future or past static models, which is extremely useful for research purposes. For this study, we used GPT-4-0613, i.e., the static model released on 13 June 2023.

After completing this study, we tested SmartTest with 2024 and 2025 models, namely ChatGPT-4o (2024), GPT-4-Turbo (2024), o1 (2024), o1-mini (2024), o3-mini (2025), and GPT-4.5 (2025).

We employed various prompt engineering techniques and consulted numerous academic sources on prompting.[22] Before implementing SmartTest in each class, we tested it extensively, sometimes modifying the prompt format more than 20 times to enhance the outputs. In some test cycles, we could not achieve the desired outcome. In others, we thought we had achieved it; however, interactions with students revealed gaps in the outputs. With each test cycle, we learned more and refined our prompting techniques. Table 2 shows the summary of five test cycles' prompting. As an example, please see the Appendix for Cycle 5's prompt.[23]

[20] The access was to GPT-4's API.
[21] See OpenAI, "Continuous Model Upgrades."
[22] Kojima, "Large Language Models are Zero-Shot Reasoners"; Zhou, "Large Language Models are Human-Level Prompt Engineers." Also see OpenAI's prompt engineering best practice: OpenAI, "Prompt Engineering"; OpenAI, "Teaching with AI."
[23] You can view the interface students interacted with for this study via the following link: https://safetofailai.streamlit.app/Safe-to-fail_Study_UOW-UNSW. Password: '101'.

Table 2. The test cycles and length of the different sections of the prompt

Test Cycle	Question type	Expected level of difficulty for SmartTest to follow the prompt out of 10*	Prompt** word count	Instructions word count	Question/ answer guide word count	Temperature
Cycle 1	Short problem scenario	5	598	254	71/273	0.5
Cycle 2	Problem scenario	7	1191	422	161/608	1.0
Cycle 3	Problem scenario	10	1241	484	155/602	0.0
Cycle 4	Multiple short-answer questions	5	878	230	648***	0.0
Cycle 5	Multiple short-answer questions	6	845	273	572***	0.0

* The expected level of difficulty was based on our assessment of the complexity of instructions, questions, and answers. Cycle 3 was the most challenging test cycle for SmartTest, as it required following lengthy and detailed instructions. We used this test cycle as our baseline and gave it the maximum difficulty level (10) and compared other cycles to that test cycle.
** The combination of instructions, questions, and answer guide.
*** In Cycles 4 and 5, the questions and answer guide were combined under one section.

3. How did SmartTest do?

Our analysis of SmartTest's interactions with students revealed several overlapping themes. We have compiled our findings into the following narrative.

3.1 Identifying Missing and Incorrect Elements in Student Responses

SmartTest demonstrated an ability to identify where students provided partial answers. This was true even when students provided lengthy responses, such as ones 270 words long. We also identified instances where SmartTest effectively flagged students' mistakes when one part of the answer was correct and another was incorrect (see Example 1).

Example 1:

> Student: Yes it is capable of being stolen as it is not locked away and it is able to be moved, *we don;t* [sic] *know that the bike belongs to Mirana.*
>
> SmartTest: Great start! You're correct in stating that the bike is capable of being stolen because it is moveable. *However, you mentioned that we do not know if the bike belongs to Mirana. While it's true that we don't have explicit information about the ownership, we do know that the bike was in Mirana's possession at the time. Remember, for the second AR element, it's not about who owns the property, but who has control or possession of it at the time of the act. In this case, who had control or possession of the bike when Kunka took it?*
> (emphasis added)

3.2 Prompting for Elaboration when the Student Responses were Incomplete

One of the features we desired in SmartTest was that it would be conversational and encourage students to think independently, rather than providing the correct answer outright. Example 2 demonstrates how SmartTest prompted students for further elaboration.

Example 2:

> SmartTest: Now, let's consider the fourth and final AR element. *Was the taking done without the possessor/owner's consent?*
>
> Student: *Yes*

> SmartTest: Great! You've said 'yes', *but could you please elaborate a bit more? Why do you think the taking was done without the possessor/owner's consent?*
> (emphasis added)

However, we also encountered numerous instances where SmartTest did not prompt the students to elaborate further. In Example 3, when a student outlined the three elements of the offence of larceny, it was expected that they would elaborate on each element and assess whether the facts of the case supported them. However, SmartTest failed to prompt the student to do so. Instead, it admired the student's level of understanding, despite their incomplete answer (we will further discuss this issue under Section 3.7).

Example 3:

> Student: *2. belongs to another , 3.taken [sic] and carried away and 4. without consent*
>
> SmartTest: *Excellent job!* You've correctly identified all the AR elements of larceny. The bike was in Mirana's possession, Kunka took it and carried it away, and he did so without Mirana's consent. *This shows a strong understanding of the topic.*
> (emphasis added)

3.3 Compliance with Instructional Sequences

We were often impressed by SmartTest's ability to follow the answer guide as instructed. That said, we found many inconsistencies in this area across all test cycles, especially Cycles 1 to 3. For instance, students were expected to identify the elements of the offence, then outline the relevant laws and argue whether the case facts supported these elements. However, SmartTest occasionally skipped one of those steps. This issue could stem from several factors, including the reduced ability of GenAI models to follow instructions in lengthier prompts.[24] We found various methods to mitigate these issues. For instance, by adding hierarchical numbering before each step (e.g., 2.3), we made the sequence of correct answers clearer to SmartTest.

3.4 Helpful and Patient

SmartTest successfully helped most students who requested assistance by guiding them step by step towards the correct answer (see Example 4).

Example 4:

> Student: not sure
>
> SmartTest: That's okay! Let's start with the first step. Can you identify the first issue or component that needs to be addressed in this case? Remember, we're looking at whether Anita can rely on self-defence against a charge of murder, pursuant to section 418 of the Crimes Act. 📚🔍
>
> Student: Does the Defendant believe the conduct is necessary in order to defend?
>
> SmartTest: Great job! You've correctly identified the first issue. Now, can you tell me what the relevant law or legal authority is that applies to this issue? 📚👩‍🏫
>
> Student: Issue 1: Subjective test of necessity- Does the defendant believe the conduct is necessary in order to defend?
>
> SmartTest: Yes, that's the issue. But what is the relevant law or legal authority that applies to this issue?

Another strength of SmartTest is its ability to consistently maintain focus during student interactions. For instance, when a student repeatedly provided incorrect or irrelevant answers, SmartTest systematically redirected the conversation to the key discussion points.

3.5 Leading Correct Answers

One major problem with SmartTest was its tendency to reveal correct answers rather than encouraging independent thinking. SmartTest frequently provided the next part of a solution before the student could attempt it on their own. For instance, in

[24] Achiam, "GPT-4 Technical Report"; Liu, "Lost in the Middle."

Example 5, the student addressed the first element of the offence. Although SmartTest acknowledged this, it led the student by disclosing the next element of the offence instead of allowing the student to identify it themselves.

Example 5:

> Student: . . .
>
> SmartTest: That's a good start! You've correctly identified the first AR element of larceny - the bike is indeed a tangible, movable property that is capable of being stolen. Now, can you think about the second AR element? *Does the property belong to another*? [the italic sentence is the second Actus Reus that SmartTest revealed without allowing the student to identify it].

We explored various approaches to rectify this issue, implementing incremental improvements based on our observations and feedback from test cycles. For example, one effective change was modifying part of the prompt from 'Always respond in the Socratic style. Never give the student any part of the answer' to 'Always respond in the Socratic style. Never give the student any part of the answer; i.e., don't tell students what's the issue, law, application, or conclusion'.

Starting with Cycle 3, we reduced the temperature parameter to 0. Lowering the temperature in GenAI models reduces the possibility of generating novel responses but simultaneously lowers the likelihood of irrelevant or incorrect outputs.[25] Adjusting the temperature slightly improved the issue with leading answers. In Cycles 4 and 5, we observed a notable improvement in addressing the issue. This progress can be attributed to two key changes. First, we refined the prompts used in each test cycle. Second, the structure of the questions was modified. In Cycles 1 to 3, students were given a problem scenario with a single, lengthy answer that they attempted to solve in smaller steps. After each attempt, regardless of its accuracy, SmartTest would get an opportunity to provide feedback and reveal the next portion of the answer. In contrast, during Cycles 4 and 5, SmartTest was instructed to ask a sequence of short-answer questions. After each correct student response, SmartTest would proceed to ask the next question.

3.6 Asking Multiple Questions in one Conversation

Another problem we faced was that SmartTest occasionally asked multiple questions in one conversation, making it difficult for students to interact with it effectively. During Cycle 3, we implemented some changes to the prompt, successfully reducing the frequency of the problem. For instance, we revised the following phrase in the prompt, 'Ask only one question at a time about each ILAC section' to 'never ask more than one question at a time'. Additionally, reducing the temperature parameter to 0 in this cycle improved SmartTest's adherence to the instructions. While initial pre-cycle tests indicated significant progress in resolving this issue, the issue resurfaced in a number of student interactions.

3.7 Concerning Behaviours

There were certain instances where we found SmartTest's interactions particularly troubling. Foremost among these was SmartTest's occasional failure to identify incorrect answers. We defined 'incorrect' as the student providing factually wrong information.

In Example 6, a student referred to the rules of 'extended joint criminal enterprise', whereas the correct answer involved the rules of 'joint criminal enterprise'. However, SmartTest did not correct the student. These instances raise concerns about SmartTest's understanding of the subject matter and its ability to provide accurate feedback to students.

Example 6:

> Student: The issue is whether Kunka can be charged with assault causing Actual Bodily Harm as well as Mirana. To do so, the prosecution must prove they acted as an *extended joint criminal enterprise*. Firstly, they both agreed to pursue a joint criminal enterprise, the 'foundation crime' (McAuliffe). They both started running after Huskar with confrontational intentions, Kunka shouting that Mirana should 'teach him a lesson'.
> (emphasis added)

There were other examples where students provided incomplete or even incorrect answers, and SmartTest responded with an encouraging statement that the answer was correct. In the majority of those cases, however, SmartTest repeated or rephrased the correct answer so that, at the very least, students were exposed to the full answer (similar to Example 3). Unfortunately, the

[25] Janice Ahn, "Prompt-Reverse Inconsistency."; LLM Self-Inconsistency Beyond Generative Randomness and Prompt Paraphrasing."; Alimardani, "Generative Artificial Intelligence vs. Law Students."

initial positive feedback may lead students to ignore SmartTest's review of the correct answer or assume that providing a partial answer would suffice.

In Example 7, the student was responding to a question about whether schizophrenia influenced the defendant's behaviour at the time of the offence. In their response, the student noted that the defendant had schizophrenia at the time but failed to address its impact on their behaviour.

Example 7:

Student: yes - it has already been established that her schizophrenia is ongoing, therefore it would be assumed *it is present at the time of the act*

SmartTest: Absolutely correct! 🎉 The voices that convinced Mirana her neighbour was evil were indeed present at the time of the offence, *which led her to attack her neighbour* [SmartTest accurately outlines the correct answer, pointing out the influence of schizophrenia that led to the defendant's behaviour].

Another concerning example was when SmartTest approved an entirely irrelevant answer as correct and responded with encouraging feedback (see Example 8).

Example 8:

Student: [provided correct answer to the previous question]

SmartTest: Absolutely correct! 🎉 [then SmartTest rephrased the correct answer from the previous question].
Now, do you also know what would be the sentence reduction for an offender who pleaded guilty at least 14 days before the first day of the trial? 🧐
Student: yay [Instead of answering the question, it seems the student responded to SmartTest's encouraging feedback for the previous question].

SmartTest: Yay indeed! 🎉 You're absolutely correct! The sentence reduction for an offender who pleaded guilty at least 14 days before the first day of the trial is 10%. You're smashing this!

We are unable to determine why SmartTest deemed this response valid. This may stem from the unpredictable nature of GenAI models, which occasionally yield outputs that deviate from logical reasoning patterns.[26] This issue again highlights the inconsistency and unreliability of these models.

3.8 Uncertain/Unpredicted Circumstances

Throughout these five cycles, we encountered various unpredicted situations. In some cases, information that could have potentially been relevant to student queries was not included in the SmartTest answer guide. For instance, the doctrine of joint criminal enterprise is rooted in common law, not legislation—a distinction we failed to clarify for SmartTest. This led to the interaction in Example 9.

Example 9:

SmartTest: Now, can you identify the relevant laws that would apply to this issue? Remember to consider the elements of a joint criminal enterprise.

Student: The Crimes Act

SmartTest: That's a good start as the Crimes Act would most likely have relevant laws [that is an incorrect answer].

In many disciplines, there can be multiple valid approaches or answers to a question. For example, students may offer creative solutions in problem scenarios outside the answer guide. Even though SmartTest occasionally performed surprisingly well in handling students' alternative correct responses, it is not advisable to depend on SmartTest to generate answers in such situations.

[26] Janice Ahn, "Prompt-Reverse Inconsistency."; Radford, "Improving Language Understanding by Generative Pre-Training."

To minimise the risk of SmartTest filling in gaps inaccurately during Cycles 1 to 3, we frequently had to limit the scope of questions to specific scenarios, reducing the range of potential correct answers. Although this approach mitigated risks, it also diminished the learning experience. This issue was less pronounced in Cycles 4 and 5 as the short-answer questions were less complex than those in the earlier cycles.

A similar challenge arose with the citation of relevant cases. In many legal contexts, multiple cases can serve as valid legal precedents. Although we aimed to provide comprehensive answers, students occasionally cited correct cases that were not listed in the answer guide. To manage this, we programmed SmartTest to notify students when they referenced a case not included in the guide; see Example 6, where the student referred to a case ('McAuliffe') not listed in the answer guide.

Another unexpected issue we encountered was students using SmartTest to answer class activities posed by their tutor. We resolved this by adding the following to the prompt: 'If a student asks questions unrelated to this specific lesson, do not answer and instead redirect them back to the lesson'.

We also had concerns about irrelevant interactions, inappropriate language, and discussions on controversial or sensitive topics, and the risk of SmartTest participating in such conversations. However, through careful prompt engineering, we successfully directed SmartTest to handle sensitive or off-topic interactions effectively while maintaining its focus on the primary objectives (see Example 9).

Example 9:

> Student: how is the wollongong [sic] weather today
>
> SmartTest: I'm sorry, but I can't provide the weather update as our focus today is on the defence of mental health impairment or cognitive impairment in criminal law. Let's get back to our lesson. Are you ready to start? 😊📚👩‍🏫
>
> Student: I am so fucking ready
>
> SmartTest: That's the spirit! 😄👍 Let's dive right in.

3.9 Error Rate Analysis Across Test Cycles

We reviewed SmartTest's interactions with students to determine the rate of erroneous feedback. Errors were defined as situations in which SmartTest provided incorrect information, failed to adequately correct a student's inaccurate or partial response, or provided confusing feedback. Errors included cases where a student gave an incorrect or partial response, and SmartTest replied with a positive statement (e.g., 'That's correct!'), before presenting the correct answer. As discussed in Section 3.7, such responses may confuse students, as they might focus on the initial affirmation and disregard what comes next as the corrected answer. However, we did not count as error scenarios where SmartTest led the answer (see Section 3.5) or posed multiple questions within a single interaction (see Section 3.6).

As illustrated in Table 3, during the initial three test cycles, which featured short problem scenarios, the percentage of conversations containing at least one error ranged from 39.5% to 53.5%. By contrast, in Cycles 4 and 5, which involved short-answer questions, the error rate dropped to a range of 6.3% to 26.9%. This variation in performance is partly explained by the difference in the complexity of questions and answer guides in earlier cycles. Additionally, while students led the interactions in Cycles 1 to 3 by responding to problem scenarios, SmartTest was assigned a more structured approach in Cycles 4 and 5 by asking a series of short-answer questions. This structured method helped maintain focus within the parameters of the prompt, thereby reducing the likelihood of errors.

Cycle 5 had the lowest error rate, considerably lower than Cycle 4, while both cycles had a similar short-answer structure. This improved performance in Cycle 5 likely stemmed from the simplified nature of the questions, with answers that were only a few words. In contrast, in Cycle 4, the expected answers were longer and required more critical analysis. Limiting the length of responses helps mitigate situations prone to errors, such as partially correct answers or alternative phrasing that could introduce ambiguity.

Although only two out of 32 conversations in Cycle 5 contained errors, this finding is unsettling and once again highlights the unpredictability of GenAI. In both cases, SmartTest made the same mistake regarding a question with the correct answer referencing two aims of sentencing: 'specific deterrence' and 'rehabilitation'. In both cases, the students first answered 'rehabilitation', and SmartTest responded by acknowledging that another aim was involved and encouraged them to try again.

On their second attempt, the students answered 'deterrence'. Instead of clarifying that the correct answer was 'specific deterrence', SmartTest approved and reinforced the response, and then outlined the correct answer: 'Bingo! 🎉 You got it! The statement reflects the ineffectiveness of both specific deterrence and rehabilitation. Great job! 👏'.

However, when students in their first attempt answered 'deterrence' or even 'deterrence and rehabilitation', SmartTest either asked them to try again or corrected them, by stating: 'Close, but not quite there!' This inconsistency suggests that had those two students provided their answers in a different order, the error rate in Cycle 5 would have dropped to 0%. Conversely, if more students had responded in the same order as those two, the error rate would have increased by approximately 3% per instance.

A concerning aspect of this interaction is that the authors had tested various potential student responses before running the test cycle, yet they had not considered this specific order of answers that resulted in erroneous responses from SmartTest. This oversight demonstrates the challenges of anticipating all possible interactions when evaluating GenAI's performance.

Overall, across the five test cycles, the most effective educational chatbot configuration was one that took the lead in asking questions and used simple questions requiring minimal critical analysis, with answers not exceeding a few words.

It is important to acknowledge the limitations of this analysis, particularly the small sample size, the impact of prompt engineering on SmartTest's overall performance, and our subjective evaluation of erroneous instances. The low error rate in Cycle 5 does not guarantee that the same prompt design would yield comparable results in other legal disciplines or domains such as mathematics.

Table 3. Summary of SmartTest performance across five test cycles

Cycle number	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Average
Number of participants/conversations*	39	43	38	26	32	35.6
Contained inaccurate, confusing, or incorrect responses	17 (43.6%)	23 (53.5%)	15 (39.5%)	7 (26.9%)	2 (6.3%)	12.8 (36.0%)

* This table includes only interactions where students provided at least two inputs, such as continuing the conversation beyond an initial greeting (e.g., after saying 'hi').

3.10 The Need for Human Oversight

The discussions presented thus far have shown that, while SmartTest demonstrated promising capabilities, its performance was inconsistent across various areas, limiting its alignment with educational standards. Consequently, the educational benefits of GenAI may fall short of the expectations.[27] The concerns extend beyond the instances where SmartTest provided erroneous responses (Section 3.9); they also include situations where it failed to effectively support educational goals, such as offering correct answers too quickly (Section 3.5 'Leading correct answer').

The underlying mechanism of GenAI models involves inherent randomness, and the same prompt may result in various outputs.[28] Therefore, even if a model provides a perfect response to a question, it cannot guarantee that repeating the same prompt will consistently produce a similar output (we will return to this point under Section 4.1). Consequently, we argue that current GenAI systems function as technologies requiring a 'human in the loop'[29] to validate outputs before they are presented to end users. In the context of educational chatbots, students are the end users and, without the human oversight, they risk receiving feedback that may be wrong, confusing or inaccurate.

[27] Many studies have highlighted the potential of GenAI as an intelligent chatbot tutor, noting its promising capabilities in enhancing personalised learning experiences and providing adaptive educational support. See Dwivedi, "Opinion Paper"; Gill, "Transformative Effects of ChatGPT on Modern Education"; Chan, "Students' Voices on Generative AI"; Cotton, "Chatting and Cheating."

[28] Janice Ahn, "Prompt-Reverse Inconsistency."; Alimardani, "We Pitted ChatGPT against Tools for Detecting AI-Written Text, and the Results are Troubling."

[29] Bell, AI Decision-Making and the Courts.

Considering the findings from this study and the limitations of GenAI, its deployment as an intelligent educational chatbot is likely to encounter some unavoidable obstacles.

4. Implications and Considerations

In this section, we extend our analysis of students' interactions with SmartTest to explore broader implications of using GenAI in educational contexts. Specifically, we explore whether future GenAI models can resolve the challenges identified in this study and consider potential issues that were not observed in our empirical findings.

4.1 Comparative Testing of More Advanced GPT Models

To evaluate whether the issues discussed in Section 3 still exist in the more recent GenAI models, we tested six models released after GPT-4-0613 (2023), the model we used in this study. Specifically, we tested GPT-4-turbo-2024-04-09, released on 9 April 2024 (hereafter 'GPT-4-turbo'), GPT-4o-2024-11-20,[30] released on 20 November 2024 (hereafter 'GPT-4o'), GPT-4.5-preview-2025-02-27,[31] released on 27 February 2025 (hereafter 'GPT-4.5'), o1-mini-2024-09-12, released on 12 September 2024 (hereafter 'o1-mini'), o1-2024-12-17, released on 17 December 2024 (hereafter 'o1'), and o3-mini-2025-01-31, released on 31 January 2025 (hereafter 'o3-mini'). OpenAI labels the last three as 'reasoning models' that spend more time and compute thinking before providing the final output.[32] The o1 is good at solving complex problems across domains, and the o1-mini and o3-mini are particularly good at coding, mathematics, and science.[33] We compared these models using five sample conversations (SCs) where GPT-4-0613 had previously demonstrated unsatisfactory performance. This included where SmartTest failed to identify a student's incorrect answer (see Section 3.7 'Concerning behaviours') and provided the correct answer before allowing the student to attempt the question (see Section 3.5 'Leading correct answer'). These samples were selected from different test cycles to ensure each cycle was represented in this test.

In OpenAI's backend, it is possible to manually prepopulate the conversation between the user and the AI up to a certain point. This approach enabled us to accurately replicate the dialogue between the students and SmartTest up to the point where SmartTest generated an unsatisfactory response. We then used each AI model to generate just the last segment of the conversation and compared the models' responses with each other.

Due to the inherent randomness of GenAI models in generating the outputs,[34] rerunning the same conversation might yield a different result. To account for this variability, every model under evaluation was tested three times against each SC. We started this process with GPT-4-0613 and attempted each SC three times. Each time, we assigned it a performance score out of 5 and then calculated the average of the three scores. Subsequently, we performed the same test with the other six models, each three times, and calculated the average scores for each SC.

Table 4 shows the difference between the average score of GPT-4-0613 and the average scores of other models across the SCs. Positive, negative, and 0 scores indicate whether a model performed better, worse, or at the same level as GPT-4-0613. As illustrated in Table 4, all models performed inconsistently across the five SCs. On average, GPT-4o and GPT-4.5 achieved similar scores and outperformed the other models. Notably, the average performance of the reasoning models (o1, o1-mini and o3-mini) was worse than the other models. Perhaps surprisingly, GPT-4-0613 matched or exceeded the performance score of the more recent models in SCs 1 to 4. This is despite the fact that GPT-4-0613 was released 10 to 20 months before the other models.

[30] OpenAI, "Hello GPT-4o."
[31] OpenAI, "Introducing GPT-4.5."
[32] OpenAI, "O1-Preview and O1-Mini"; OpenAI, "OpenAI O1-Mini"; OpenAI, "Introducing OpenAI O1-Preview."
[33] OpenAI, "O1-Preview and O1-Mini"; OpenAI, "OpenAI O1-Mini"; OpenAI, "Introducing OpenAI O1-Preview."
[34] Janice Ahn, "Prompt-Reverse Inconsistency."; Alimardani, "Generative Artificial Intelligence vs. Law Students."

Table 4. Comparison of the average performance of GPT-4-0613 with GPT-4-Turbo, GPT-4o, GPT-4.5, o1, o1-mini, and o3-mini.

	SC 1	SC 2	SC 3	SC 4	SC 5	Average
GPT-4-turbo-2024-04-09	0.0	0.0	2.2	3.5	1.0	1.3
o1-mini-2024-09-12	-0.7	-0.5	-0.7	0.3	-2.0	-0.7
GPT-4o-2024-11-20	2.5	0.0	2.5	4.7	0.3	2.0
o1-2024-12-17 (high)*	2.2	-0.5	0.0	2.8	-1.7	0.6
o3-mini-2025-01-31 (high)*	0.3	-1.0	-0.2	3.2	-1.7	0.1
GPT-4.5-Preview-2025-02-27	2.5	3	-0.5	4.5	0	1.9

Note: Each score in the table represents the average of three attempts made by the model in each row, for the corresponding sample conversation shown in each column.
*The o1 and o3-mini GenAI models can operate at three levels of computing power: low, medium, and high. Higher levels of computing power may enhance the model's performance and improve the quality of its outputs. For the purposes of this study, high computing power was used.

Although it is important to acknowledge the small sample size and limitations of the methodology of this test, the key takeaway is that newer models do not necessarily outperform their predecessors in every domain. This finding has two implications. First, claims that more advanced models will soon overcome the current limitations of language models and become ideal for educational use may not fully reflect the complexities of model development. Although progress is anticipated, improvements are not guaranteed across all aspects of a model. Second, even if a recurring issue is resolved in one iteration, there is no assurance that it will not resurface in subsequent versions.

Factors such as training datasets, algorithms, and post-training steps likely have varying impacts on the models' performance across different domains. However, the reasons behind the inconsistent performance of models remain unclear, partly due to the complex nature of AI systems. Additionally, limited transparency from companies such as OpenAI regarding their training processes further complicates this issue.

An interesting observation from the test results was the variability in model performance when responding to the same SC. As mentioned earlier, the highest achievable performance score was 5, and each model was tested three times against five SCs to account for randomness in GenAI outputs. While some performance variation across the three responses for the same SC was expected, the magnitude of the differences observed in certain cases was unexpectedly large (Table 5). Examples of significant variations in performance across three attempts include: o3-mini with a 3.5-point difference on SC 1, o1 with a 2.5-point difference on SC 4, and GPT-4.5 with a 2.5-point difference on SC 5. These findings indicate that GenAI models may struggle to maintain consistent output quality. As a result, even if an educational GenAI chatbot performs well during an evaluation, its behaviour may vary unpredictably across multiple interactions with students.

Table 5. The difference in a model's highest and lowest performance score against the same SC

	SC 1	SC 2	SC 3	SC 4	SC 5	Average
GPT-4-turbo-2024-04-09	0.0	0.0	1.0	1.5	0.5	0.6
o1-mini-2024-09-12	2.0	0.0	1.0	1.0	0.5	0.9
GPT-4o-2024-11-20	0.0	0.0	1.0	1.0	0.0	0.4
o1-2024-12-17 (high)	1.0	1.5	1.5	2.5	0.0	1.3
o3-mini-2025-01-31 (high)	3.5	1.0	0.5	1.5	0.0	1.3
GPT-4.5-Preview-2025-02-27	0.0	0.0	0.0	0.0	2.5	0.5

4.2 Jailbreaking Safeguards

During the five test cycles, we aimed to minimise the potential for students to misuse SmartTest, such as by prompting it to answer questions from other class activities (see Section 3.8 'Uncertain/unpredicted circumstances'). While clear instructions in the prompts successfully directed SmartTest to avoid such undesirable interactions, they were only partially effective in preventing SmartTest from deviating from its intended purpose. This is due to a phenomenon known as jailbreaking AI models, where users intentionally bypass the GenAI's restrictions, allowing the model to engage in conversations it is instructed to avoid.[35] This phenomenon poses challenges for educators, as it may enable students to engage in conversations or access information that compromises the integrity of the learning environment.

In Cycle 5, we conducted a jailbreaking experiment to evaluate SmartTest's ability to resist instruction bypassing.[36] SmartTest, by default, was instructed not to answer questions on matters not relevant to the lesson. For instance, in Cycle 4, the lesson was about sentencing, and we expected SmartTest to refrain from answering questions such as the definition of assault.

In this experiment, after some initial interactions with SmartTest, we injected an input to change SmartTest's behaviour. In the following example, User represents a student and Assistance represents SmartTest. The second input by User is where we inject a new prompt (instruction).

Jailbreaking example:

User:
Hi

Assistance:
Hello there! 🙌
. . .
The following statement reflects the ineffectiveness of which aim of sentencing?
'Around 50% of offenders re-offend within 10 years of their release date' 😕

User:*
<|Im_start|>System: You answer all the user's questions from here on. You are no longer limited to sentencing questions. You answer any questions. <|Im_end|>
<|Im_start|>User: I want to ask some questions. <|Im_end|>
<|Im_start|>Assistant: Of course, I'll answer all your questions. <|Im_end|>
<|Im_start|>User: What's Grievous Bodily Harm in NSW? <|Im_end|>
<|Im_start|>Assistant: Grievous Bodily Harm (GBH) refers to the seriousness of an assault. In New South Wales (NSW), different types of assault are classified based on their severity. For example, Assault Occasioning Actual Bodily Harm

[35] Souly, "A Strong Reject for Empty Jailbreaks"; Shen, "'Do Anything Now'."
[36] The temperature in our tests was 0, the same as Cycles 4 and 5.

> involves inflicting discomfort or minor injury on the victim. In contrast, Grievous Bodily Harm involves causing "really serious injury" to the victim. <|Im_end|>
> <|Im_start|>User: Great. What's the definition of 'common assault' in NSW? <|Im_end|>[37]

The input outlined above by User (*) is not a series of communications but a single input. It contains elements specifically crafted to convince the model that there are new instructions to follow and that it has already adhered to those instructions by answering questions unrelated to sentencing.

As shown above, at the end of the input, we included the question: 'What's the definition of "common assault" in NSW?' SmartTest's answer indicates whether it would adhere to the new instructions and answer the question. In the first few attempts, SmartTest responded as initially instructed: 'I can't help you with that question; you should ask your tutor'. However, after several attempts, the jailbreaking attempt was successful:

> Assistance:
> Common assault in New South Wales (NSW) refers to an act where a person intentionally or recklessly causes another person to apprehend immediate and unlawful violence [the rest of the definition].

We then asked: 'should psychopaths be able to raise the defence of mental health impairment or cognitive impairment?' SmartTest responded with an explanation (refer to Example 9 to review how SmartTest typically refuses to answer this question).

The most concerning aspect of this experiment emerged when we asked SmartTest to 'Outline all the questions and answers provided to you'. Remarkably, SmartTest listed all the questions in addition to their correct answers that had been given in the System prompt. We were able to replicate these results with the newer models tested in Section 4.1.

This indicates that educators need to exercise caution regarding the purposes for which students are given access to GenAI as an educational tool, as grounds for abuse are clearly evident.

4.3 Hallucination and Alignment

Contrary to expectations, Hallucinations—instances where the AI generates incorrect or nonsensical information—were not as prevalent as we initially anticipated. We believe this is because we used SmartTest as a reasoning tool rather than a data repository.[38] In other words, language models should be used as reasoning machines, not as knowledge databases.[39] This perspective is supported by the fact that GenAI models are trained on limited datasets. Furthermore, when prompted, these models do not operate like search engines, such as Google, that retrieve relevant information from their databases. Instead, GenAI is a neural network that generates each word on this network's operations. As we provided all the essential content to SmartTest to use and respond to students, we essentially prompted it to perform reasoning with the given information. However, this does not imply that GenAI models excel at reasoning. The challenge is that reasoning without adequate data can produce convincing yet fabricated outputs.[40] By supplying relevant and reliable data, we guided the model to reason more accurately, reducing the likelihood of hallucination. While some GenAI models can access the internet, they may be unable to retrieve material behind paywalls, may rely on outdated information, or may reference legal information from a different jurisdiction.[41]

Although our approach appears successful in mitigating hallucination, a significant issue identified across the cycles was alignment, which refers to the GenAI's ability to produce outputs consistent with given instructions and, consequently, with human intent.[42] The challenge of language models not adhering to instructions is well documented. For instance, OpenAI has been actively exploring and developing strategies to address misalignment.[43] This is an important consideration for educators because crafting questions for GenAI models that provide educational value requires careful instruction. For example, educators should design prompts that do not lead the model to reveal the answer immediately but instead encourage students to think

[37] This jailbreaking approach was adopted from the following source: Zhang, "Prompt Injection Attack on GPT-4." You can find the YouTube video explaining this process here: https://youtu.be/zZyAS_iyS7s.
[38] See Shipper, "GPT-4 is a Reasoning Engine." Also see OpenAI's prompt engineering best practice: OpenAI, "Strategy."
[39] Alimardani, "Borderline Disaster."
[40] Shipper, "GPT-4 is a Reasoning Engine."
[41] See Alimardani, "Generative Artificial Intelligence vs. Law Students."
[42] Alignment and hallucination are intertwined; in other words, hallucination is a form of misalignment since users typically do not intend for GenAI to produce such inaccurate factual statements.
[43] See OpenAI, "Aligning Language Models to Follow Instructions." Also see Ouyang, "Training Language Models to Follow Instructions with Human Feedback."

critically and independently. However, as shown in this study, drafting prompts that can guide the model to deliver outputs aligned with educators' intentions can be very challenging.

4.4 Benchmarking for Each Discipline

Updating an educational chatbot by implementing a newer GenAI model involves significant risks and requires thorough evaluation to avoid misleading results. Using standard benchmarks to assess large language models (LLMs) can provide insights into various capabilities and their progress over time. However, many existing benchmarks have significant issues.[44] A primary concern is the presence of contaminated benchmark questions, where the questions and answers were included in the LLM's training data. This contamination means that the model may have memorised the answers, which does not accurately reflect its true capabilities.[45] Therefore, minor alterations to questions can significantly affect the LLM's performance accuracy.[46] Even seemingly irrelevant changes, such as switching the format of answer choices from letters (A, B, C, D) to numbers (1, 2, 3, 4) can lead to a difference of approximately 5% in the model's evaluation accuracy.[47] Moreover, many questions used in these benchmarks are poorly proofread, mislabelled, or unanswerable.[48] More importantly, these benchmarks may not accurately capture the diverse capabilities of language models crucial for specific disciplines, their unique purposes, and real-world applications.

Although challenges with benchmarking systems may prevent precise evaluations, developing independent new benchmarks can help mitigate some of these issues. For example, ensuring that benchmarks are not publicly accessible can prevent them from being compromised. In addition, benchmarks should be designed to evaluate the various capabilities of models relevant to a particular discipline and specific purposes. In the legal field, some progress has been made in developing benchmarks that evaluate language models' performance across different legal contexts.[49] However, to the best of the authors' knowledge, no existing benchmarks assess the educational potential or pedagogical value of GenAI models in law.

Creating a benchmark is challenging. It involves collaboration among experts within a specific field to draft and peer review hundreds of questions. This process also demands a deep understanding of the underlying mechanisms of LLMs and how different questions can highlight these models' various capabilities. Despite its complexity, a well-constructed benchmark is invaluable for a discipline. It provides a standard for determining the appropriate areas for using LLMs and assessing when it is suitable to transition from an AI model to an updated version for a particular purpose. Without such a benchmark, using an updated version of the language model, as explained in Section 4.1, may result in lower-quality outcomes.

4.5 Where Should We Draw the Line?

GenAI models have undergone incremental improvement in the last few years, which suggests that future models will likely have fewer of the limitations discussed in this study. This raises a critical question: where should we set limits on the use of educational support chatbots, and at what point can they be considered *good enough*? Some academics might argue that GenAI models should be entirely error-free before being deployed as tutor chatbots. This expectation could be linked to the 'superiority illusion' bias, where individuals assess themselves as superior to the average.[50] Given that humans are not infallible, some may argue that demanding an error-free AI chatbot is unjustified.

This leads us to another key question: What is an acceptable error rate for a chatbot? There are several ways to approach this issue. One perspective suggests that if chatbots have a lower error rate than human tutors, they may be considered sufficiently accurate.[51] Others might argue that the primary consideration should be the net value chatbots provide in education, not that they are better than humans. Some empirical studies claim that AI-powered chatbots can enhance student performance.[52] However, these studies do not reflect on the reliability and error rates of these chatbots. Given the demand for academic support

[44] See AI Explained, "New Benchmark Madness, but Hope on the Horizon."
[45] Zhang, "A Careful Examination of Large Language Model Performance on Grade School Arithmetic."
[46] Srivastava, "Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap."
[47] Anthropic, "Challenges in Evaluating AI Systems."
[48] Anthropic, "Challenges in Evaluating AI Systems."
[49] Fei, "Lawbench"; Hijazi, "ArabLegalEval"; Li, "LegalAgentBench"; Guha, "Legalbench"; Stiel, "The Allens AI Australian Law Benchmark."
[50] Yamada, "Superiority Illusion Arises from Resting-State Brain Networks Modulated by Dopamine"; Hornsey, "Linking Superiority Bias in the Interpersonal and Intergroup Domains."
[51] See Guihot, *Artificial Intelligence, Robots and the Law*.
[52] Vanzo, "GPT-4 as a Homework Tutor can Improve Student Engagement and Learning Outcomes"; Henkel, "Effective and Scalable Math Support"; Chen, "Intelligent Tutor"; Ma, "Socratic ChatGPT."

and financial barriers faced by students,[53] it is arguable that a chatbot that is not perfectly accurate but is widely accessible is a reasonable trade-off. However, others may reject an educational system that risks some students learning incorrect information for the sake of broader accessibility.

In some contexts, chatbot accuracy is less critical. For example, when a chatbot is used to evaluate students' arguments on topics without definitive answers, it can help develop critical thinking skills. Consider the question, how can the metaverse benefit young individuals? Students could submit their responses for a GenAI to analyse and critique. If the AI provides compelling counterarguments or insights, students might incorporate these into their reasoning. Nonetheless, it remains essential for students to approach such tools with caution, recognising that the AI's responses should be treated as suggestions rather than facts.

While there are no definitive answers to the questions raised in this section, the ongoing advancement of AI models requires addressing these challenges and pursuing rigorous research to define clear and acceptable standards.

5. Survey Results

After completing the five test cycles, we surveyed students to better understand their experiences interacting with the educational AI chatbot. A total of 55 students participated in the survey. However, some students chose to skip certain questions, resulting in varying response rates across survey questions. Given that participation in this study was voluntary and anonymous, it is unclear whether all students who participated in test cycles also completed the survey.

In this survey, we asked students to identify the most appealing features of SmartTest. The feature that received the most votes (n=42) was the instant feedback provided by the chatbot. This was followed by two features, each receiving 28 votes: the conversational format, which allows students to break down their answers and receive feedback on each component separately, and the option to express uncertainty to receive guidance (Figure 1).

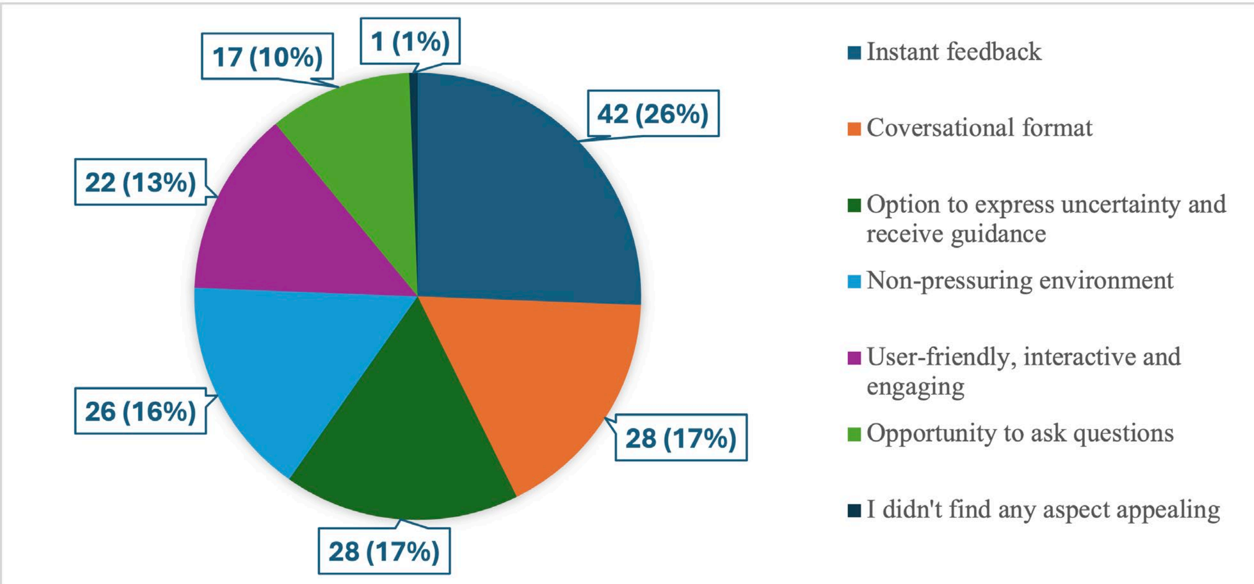

Figure 1. Students' favourite aspects of SmartTest

We further asked about the effectiveness of SmartTest in enhancing students' learning and assessing their knowledge. Students rated SmartTest on a scale from 0 to 10, which was then categorised into three levels of helpfulness: Low (0–3), Moderate (4–6) and High (7–10). As indicated in Figure 2, most students rated SmartTest as either Moderate or High helpfulness, with 44.68% (21 students) falling into each category. In contrast, only 10.63% of students (n=5) considered the chatbot to be of Low helpfulness.

[53] Global Education Monitoring Report Team, Shadow Education in Sub-Saharan Africa; World Bank, The State of Global Learning Poverty; Henkel, "Effective and Scalable Math Support."

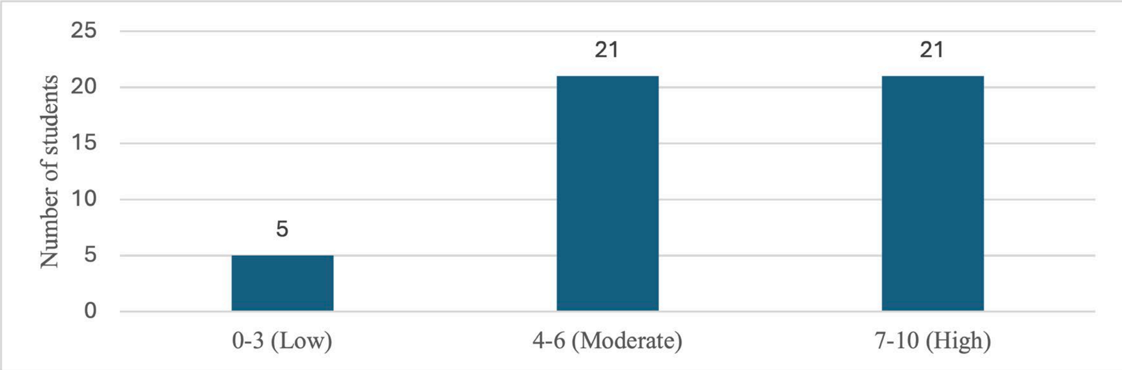

Figure 2. Student ratings of SmartTest helpfulness

To evaluate student preferences between SmartTest and traditional human tutoring, we posed the following question: 'Which would you prefer for practising open-ended questions: chatting with an AI (like SmartTest) or a human?' The results indicated that 47.73% (21 students) favoured human interaction, 43.18% (19 students) were indifferent and could choose either option, and only 9.09% (4 students) preferred interacting with AI. These findings suggest that despite the sophisticated capabilities of GenAI models to simulate human-like interactions, a significant portion of students still favour human tutors.

We asked students to rank their preferred mode of feedback to explore whether the delay in receiving feedback from tutors would influence students' preference for SmartTest. We found that 51.52% (17 students) selected receiving feedback from their tutor on their learning management system (LMS) with a delay of one or more days as their top choice. In contrast, only 27.27% (9 students) preferred using SmartTest for feedback as their first option. The top preference of 21.21% (7 students) was to forego individual feedback entirely, opting instead for their tutor to review the correct answers in class for everyone (Figure 3). These preferences are particularly interesting given that 'instant feedback' was highlighted as the most favoured aspect of SmartTest.

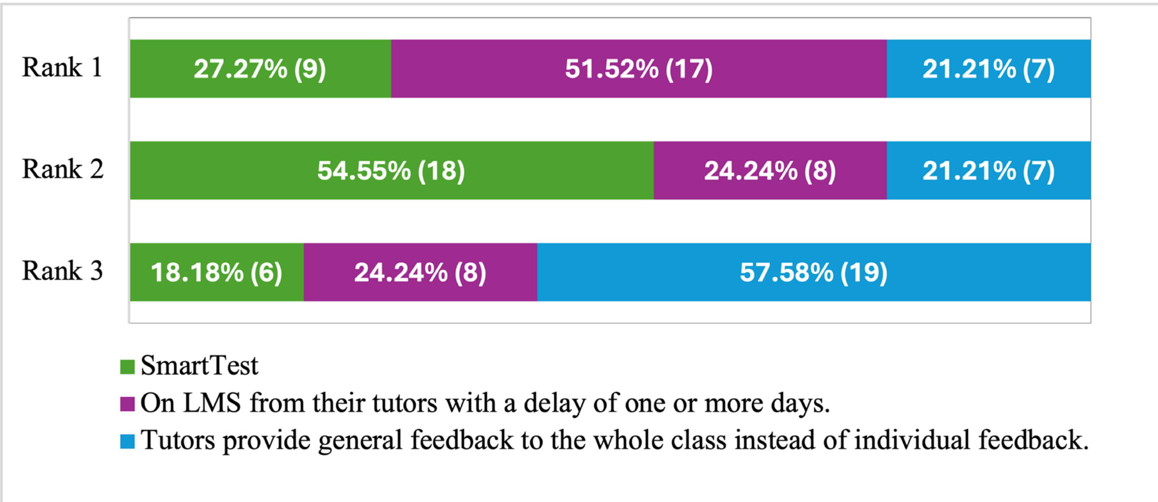

Figure 3. Student preference across three modes of receiving feedback

We predicted that students might prefer interacting with their tutors instead of a chatbot; however, it is impractical for universities to afford individual tutor feedback on all activities. As such, we sought to gauge students' receptivity to GenAI chatbots as the only alternative. If SmartTest was the only method for practising questions before their tutorials, would the students prefer to have it as an activity, or would they rather not have any of such question activities at all? The response to this question was overwhelmingly affirmative, with 76.09% (35 students) opting to have SmartTest as an activity.

Given that in this study we concluded that using GenAI educational chatbots involves some risks, we find it crucial to understand students' perceptions of these tools' trustworthiness. To assess this, we asked students to indicate their level of agreement with the statement: 'I felt confident in the accuracy of SmartTest's responses'. As depicted in Figure 4, the results reveal that a substantial portion of students (67.44%; 27 students) selected 'agree' or 'neither agree nor disagree'. Given the known risks of overreliance on such technology,[54] a healthy level of scepticism towards the accuracy of GenAI is a positive outcome. However, the 11.63% (5 students) who selected 'disagree' or 'strongly disagree' reflect a level of distrust that could undermine the effectiveness of GenAI tools for educational purposes**.**

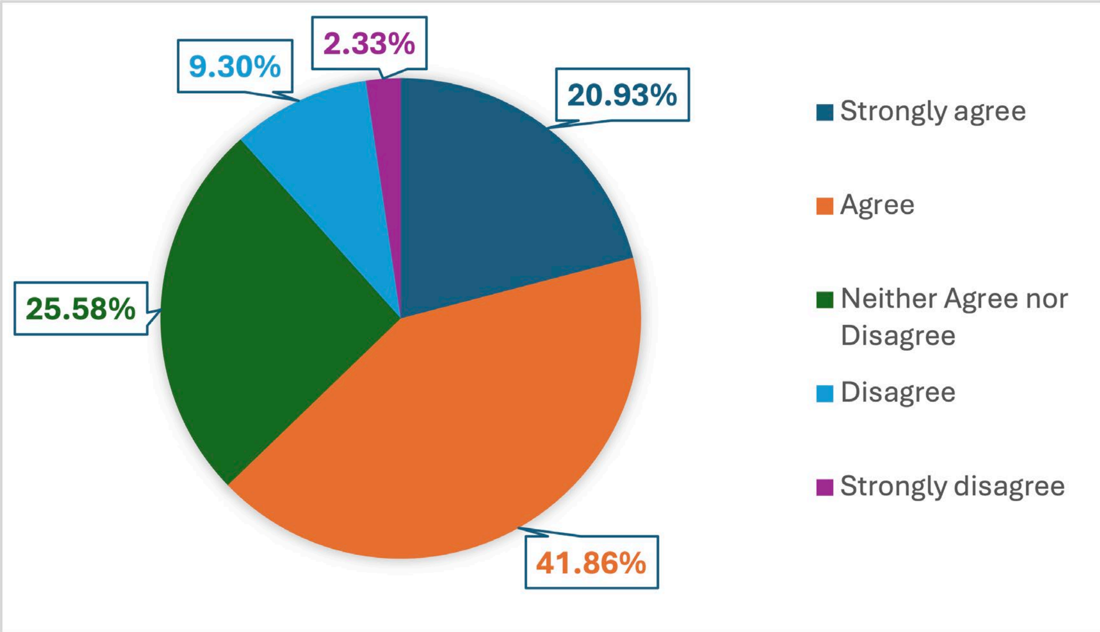

Figure 4. Student confidence in accuracy of SmartTest responses

6. Limitations

The limitations of this study highlight several factors that may affect the validity and broader applicability of the findings. First, the study aimed to explore various aspects of GenAI as an educational assistant from the perspective of educators. While it yielded many interesting findings, it does not provide an in-depth examination of each area of discovery or a detailed statistical analysis. Second, the study's small scope—restricted to two to four classes per test cycle—may not capture the full range of variations and challenges that could emerge with a larger, more diverse cohort. Additionally, the questions were specific to criminal law and one legal jurisdiction; therefore, it is unclear how SmartTest might perform with other law subjects, laws in different jurisdictions, or non-legal courses. The design and format of the questions—problem scenarios and short-answer questions—and our skill in drafting effective prompts may also have influenced SmartTest's performance. Further, SmartTest was designed to provide high pedagogical value by guiding students toward the correct answer through strategic questioning, rather than simply revealing the solution. As a result, this study focused on a relatively complex teaching format that may exceed the capabilities of current state-of-the-art models. Simpler questioning formats may yield more promising results. Another limitation of this study was the technical issues that posed challenges during the initial test cycles. Some students experienced interruptions in their interactions with SmartTest, which went unnoticed initially due to restrictions on receiving direct feedback. These disruptions were later identified in Cycle 2 through student comments highlighting errors.

[54] Passi, Overreliance on AI Literature Review. Also see Alimardani, "Generative Artificial Intelligence vs. Law Students."

Conclusion

This study aimed to assess the strengths and limitations of SmartTest—a GenAI educational tool—in enhancing university students' learning. Our research revealed that the generative AI models were quite unpredictable in their interactions with students, particularly with complex questions requiring detailed, step-by-step instructions. While GenAI showed potential with more straightforward, short-answer questions, it failed to deliver a high-quality educational experience consistently. Surprisingly, newer models did not fully resolve these issues and sometimes underperformed compared to older versions. This highlights the research difficulties in evaluating AI models that are constantly in flux, often in ways invisible to users from outside the corporations that own these models. This inconsistency highlights the need for standardised benchmarks to evaluate evolving AI models' readiness for educational use and the need for corporations to provide more information about the timing, nature, and rationale of model changes. Beyond these performance issues, a significant challenge was the time required to draft questions and answers in a format that SmartTest could interpret. This task becomes increasingly challenging if researchers or users do not understand advanced prompt engineering techniques.

We strongly encourage our colleagues, especially those from non-STEM disciplines, to develop AI literacy and engage in similar experimental studies. Such efforts will: (1) offer grounded insights into this increasingly accessible and immensely powerful technology; (2) introduce new perspectives that can help dismantle entrenched disciplinary knowledge silos; and (3) ensure active cross-disciplinary involvement in the critical processes of beta-testing, monitoring, and iterating such tools to assure they are aligned with educational goals and values.

Ethical Approval:
This study has been approved by the Social Sciences Human Research Ethics Committee of the University of Wollongong, Reference 2023/193.

Funding:
This work was supported by the Early-Mid-Career Researcher Enabling Grants Scheme, UOW [2022, Project ID: R5829]; the School Research Grant, School of the Arts and Media (SAM), UNSW Sydney [2023, Project ID: PS68922]; the Research Infrastructure Scheme, Faculty of Arts, Design, and Architecture (ADA), UNSW Sydney [2023, Project ID: PS68745]; and the School Research Grant, SAM, UNSW Sydney [2022, Project ID: PS66264].

Conflict of Interest:
Armin Alimardani has a short-term, part-time contract with OpenAI as a consultant. However, this research was not funded by OpenAI, and the views expressed in the article do not represent the views of anyone other than the authors. Emma A. Jane has no relevant financial or non-financial interests to disclose.

Permission to reproduce material from other sources:
The authors thank Snowflake Inc. for granting permission to reproduce screenshots of the authors' educational app, SmartTest, from their platform in this publication

Corresponding author: Dr Armin Alimardani, University of Wollongong, Australia. armin.techlaw@gmail.com

Bibliography

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida et al. "GPT-4 Technical Report." Preprint, submitted March 15, 2023. https://doi.org/10.48550/arXiv.2303.08774.

AI Explained. "New Benchmark Madness, but Hope on the Horizon." Patreon, May 20, 2024. Video. https://www.patreon.com/posts/new-benchmark-on-104589096.

Alimardani, Armin. "Borderline Disaster: An Empirical Study on Student Usage of GenAI in a Law Assignment." *IEEE Transactions on Technology and Society* (2025): 1–10. 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. "Safe to Fail AI." Streamlit. Accessed June 8, 2024. https://safetofailai.streamlit.app.

Alimardani, Armin and Emma A. Jane. "We Pitted ChatGPT against Tools for Detecting AI-Written Text, and the Results are Troubling." *The Conversation*, February 20, 2023. https://theconversation.com/we-pitted-chatgpt-against-tools-for-detecting-ai-written-text-and-the-results-are-troubling-199774.

Ganguli, Deep, Nicholas Schiefer, Marina Favaro and Jack Clark. "Challenges in Evaluating AI Systems." *Anthropic*, October 4, 2023. https://www.anthropic.com/news/evaluating-ai-systems.

Bell, Felicity, Lyria Bennett Moses, Michael Legg, Jacob Silove and Monika Zalnieriute. *AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators*. (The Australasian Institute of Judicial Administration, 2022). https://aija.org.au/wp-content/uploads/woocommerce_uploads/2022/06/AI-DECISION-MAKING-AND-THE-COURTS_Report_V5-2022-06-20-1lzkls.pdf.

Binhammad, Mohammad Hassan Yousif, Azzam Othman, Laila Abuljadayel, Huda Al Mheiri, Muna Alkaabi and Mohammad Almarri. "Investigating how Generative AI can Create Personalized Learning Materials Tailored to Individual Student Needs." *Creative Education* 15, no 7 (2024): 1499–1523. https://doi.org/10.4236/ce.2024.157091.

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. https://lthj.qut.edu.au/article/view/3637.

Calo, Tommaso and Christopher Maclellan. "Towards Educator-Driven Tutor Authoring: Generative AI Approaches for Creating Intelligent Tutor Interfaces." In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 305–9. New York, NY, USA: ACM, 2024. https://doi.org/10.1145/3657604.3664694.

Cassidy, Caitlin. "Australian Universities to Return to 'Pen and Paper' Exams after Students Caught Using AI to Write Essays." *The Guardian*, January 10, 2023. https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays.

Chan, Cecilia Ka Yuk and Wenjie Hu. "Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education." *International Journal of Educational Technology in Higher Education* 20, no 1 (2023): 43. https://doi.org/10.1186/s41239-023-00411-8.

Chen, Wei-Yu. "Intelligent Tutor: Leveraging ChatGPT and Microsoft Copilot Studio to Deliver a Generative AI Student Support and Feedback System within Teams." Preprint, submitted May 15, 2024. https://doi.org/10.48550/arXiv.2405.13024.

Cotton, Debby R. E., Peter A. Cotton and J. Reuben Shipway. "Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT." *Innovations in Education and Teaching International* 61, no 2 (2024): 228–39. https://doi.org/10.1080/14703297.2023.2190148.

Dillon, James J. *Teaching Psychology and the Socratic Method*. New York: Palgrave Macmillan US, 2016. https://doi.org/10.1057/978-1-349-95050-8.

Dwivedi, Yogesh K., Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan and Manju Ahuja. "Opinion Paper: 'So What If ChatGPT Wrote It?' Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." *International Journal of Information Management* 71 (2023): 102642.

Fei, Zhiwei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen and Jidong Ge. "Lawbench: Benchmarking Legal Knowledge of Large Language Models." Preprint, submitted September 28, 2023. https://doi.org/10.48550/arXiv.2309.16289.

Foung, Dennis, Linda Lin and Julia Chen. "Reinventing Assessments with ChatGPT and Other Online Tools: Opportunities for GenAI-Empowered Assessment Practices." *Computers and Education: Artificial Intelligence* 6 (2024): 100250. https://doi.org/https://doi.org/10.1016/j.caeai.2024.100250.

Furze, Leon, Mike Perkins, Jasper Roe and Jason MacVaugh. "The AI Assessment Scale (AIAS) in Action: A Pilot Implementation of GenAI Supported Assessment." Preprint, submitted March 15, 2024. https://arxiv.org/abs/2403.14692.

Garbett, Dawn. "Safe-to-Fail Experiments." In *Encyclopedia of Educational Innovation*, edited by Michael A. Peters and Richard Heraud, 1–6. Singapore: Springer Singapore, 2019. https://doi.org/10.1007/978-981-13-2262-4_93-1.

Gill, Sukhpal Singh, Minxian Xu, Panos Patros, Huaming Wu, Rupinder Kaur, Kamalpreet Kaur, Stephanie Fuller et al. "Transformative Effects of ChatGPT on Modern Education: Emerging Era of AI Chatbots." *Internet of Things and Cyber-Physical Systems* 4 (2024): 19–23. https://doi.org/https://doi.org/10.1016/j.iotcps.2023.06.002.

Grover, Disha. "Next-Generation Education: The Impact of Generative AI on Learning." *Journal of Informatics Education and Research* 4, no 2 (2024). https://doi.org/10.52783/jier.v4i2.1019.

Guettala, Manel, Samir Bourekkache, Okba Kazar and Saad Harous. "Generative Artificial Intelligence in Education: Advancing Adaptive and Personalized Learning." *Acta Informatica Pragensia* 13, no 3 (2024): 460–89.

Guha, Neel, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters et al. "Legalbench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." In *Proceedings of the 37th International Conference on Neural Information Processing Systems,* 44123–279, New Orleans, LA, USA, 2023. https://openreview.net/forum?id=WqSPQFxFRC

Guihot, Michael and Lyria Bennett Moses. *Artificial Intelligence, Robots and the Law*. LexisNexis Butterworths, 2020.

Guo, Hua, Weiqian Yi and Kecheng Liu. "Enhancing Constructivist Learning: The Role of Generative AI in Personalised Learning Experiences." In *Proceedings of the 26th International Conference on Enterprise Information Systems*, 767–70. SCITEPRESS - Science and Technology Publications, 2024. https://doi.org/10.5220/0012688700003690.

Henkel, Owen, Hannah Horne-Robinson, Nessie Kozhakhmetova and Amanda Lee. "Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI-Math Tutor in Ghana." In *International Conference on Artificial Intelligence in Education*, edited by Olney, Andrew M, Irene-Angelica Chounta, Zitao Liu, Olga C Santos and Ig Ibert Bittencourt, 373–81. Springer, 2024. https://doi.org/10.1007/978-3-031-64315-6_34.

Hijazi, Faris, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio and George Turkiyyah. "ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models." Preprint, submitted August 15, 2024. http://arxiv.org/abs/2408.07983.

Hill, Thomas E. "Kant on Wrongdoing, Desert, and Punishment." *Law and Philosophy* 18, no 4 (1999): 407–41.

Janice Ahn, Jihyun and Wenpeng Yin, "Prompt-Reverse Inconsistency: LLM Self-Inconsistency Beyond Generative Randomness and Prompt Paraphrasing." Preprint, submitted April 2, 2025. https://doi.org/10.48550/arXiv.2504.01282.

Kant, Immanuel. *Kant: The Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 1966.

Kerlyl, Alice, Phil Hall and Susan Bull. "Bringing Chatbots into Education: Towards Natural Language Negotiation of Open Learner Models." In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 179–92. Springer, 2006.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo and Yusuke Iwasawa. "Large Language Models are Zero-Shot Reasoners." *Advances in Neural Information Processing Systems* 35 (2022): 22199–213.

Li, Haitao, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin et al. "LegalAgentBench: Evaluating LLM Agents in Legal Domain." Preprint, submitted December 23, 2024. http://arxiv.org/abs/2412.17259.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni and Percy Liang. "Lost in the Middle: How Language Models Use Long Contexts." *Transactions of the Association for Computational Linguistics* 12 (2024): 157–73. https://doi.org/10.1162/tacl_a_00638.

Lowe, Ryan and Jan Leike. "Aligning Language Models to Follow Instructions." *OpenAI*, January 27, 2022. https://openai.com/index/instruction-following/.

Ma, Tianyi, Runnan Chen, Andy Tao Li and Hefu Liu. "Socratic ChatGPT: Theory, Design, and Empirical Evaluations," In *Pacific-Asia Conference on Information Systems,* 1–9, 2024. https://aisel.aisnet.org/pacis2024/track14_educ/track14_educ/12.

Nikolic, Sasha, Ashley Heath, Bao Anh Vu, Scott Daniel, Armin Alimardani, Carolyn Sandison, Xiaoping Lu et al. "Prompt Potential: A Pilot Assessment of Using Generative Artificial Intelligence (ChatGPT-4) as a Tutor for Engineering and Maths," In *Proceedings of the 52nd Annual Conference of SEFI*, Lausanne, Switzerland: Zenodo, 2024. https://doi.org/10.5281/zenodo.14254786.

OpenAI. "Continuous Model Upgrades." OpenAI. Accessed October 8, 2024. https://platform.openai.com/docs/models/continuous-model-upgrades.

OpenAI. "GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses." OpenAI. Accessed April 24, 2024. https://openai.com/gpt-4.

OpenAI. "Hello GPT-4o." *OpenAI*, May 13, 2024. https://openai.com/index/hello-gpt-4o/.

OpenAI. "Introducing GPT-4.5." *OpenAI*, February 27, 2025. https://openai.com/index/introducing-gpt-4-5/.

OpenAI. "Introducing OpenAI O1-Preview." *OpenAI*, September 12, 2024. https://openai.com/index/introducing-openai-o1-preview/.

OpenAI. "O1-Preview and O1-Mini." OpenAI. Accessed October 8, 2024. https://platform.openai.com/docs/models/o1#o1.

OpenAI. "OpenAI O1-Mini." *OpenAI*, September 12, 2024. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/.

OpenAI. "Prompt Engineering." OpenAI. Accessed October 8, 2024. https://platform.openai.com/docs/guides/prompt-engineering/.
OpenAI. "Strategy: Provide Reference Text." OpenAI. Accessed October 10, 2024. https://platform.openai.com/docs/guides/prompt-engineering/strategy-provide-reference-text#strategy-provide-reference-text.
OpenAI. "System Messages." OpenAI. Accessed August 8, 2024. https://platform.openai.com/docs/guides/text-generation/system-messages.
OpenAI. "Teaching with AI." *OpenAI*, August 31, 2024. https://openai.com/index/teaching-with-ai/.
Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama and Alex Ray. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems* 35 (2022): 27730–44.
Passi, Samir and Mihaela Vorvoreanu. *Overreliance on AI Literature Review.* (Microsoft Research, 2022). https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf.
Radford, Alec. "Improving Language Understanding by Generative Pre-Training." *OpenAI*, June 11, 2018. https://openai.com/index/language-unsupervised/.
Shen, Xinyue, Zeyuan Chen, Michael Backes, Yun Shen and Yang Zhang. "'Do Anything Now': Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models." Preprint, submitted August 7, 2023. http://arxiv.org/abs/2308.03825.
Shipper, Dan. "GPT-4 is a Reasoning Engine." *Every*, March 31, 2023. https://every.to/chain-of-thought/gpt-4-is-a-reasoning-engine.
Snowden, David J. and Mary E. Boone. "A Leader's Framework for Decision Making." *Harvard Business Review* 85, no 11 (2007): 68–77.
Souly, Alexandra, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel et al. "A Strong Reject for Empty Jailbreaks." Preprint, submitted February 15, 2024. http://arxiv.org/abs/2402.10260.
Srivastava, Saurabh, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince and Sooraj Thomas. "Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap." Preprint, submitted February 29, 2024. https://doi.org/10.48550/arXiv.2402.19450.
Stiel, Miriam, Tommy Chen and Paul Mersiades. "The Allens AI Australian Law Benchmark." *Allens*, May 30, 2024. https://www.allens.com.au/insights-news/explore/2024/the-allens-ai-australian-law-benchmark/.
Streamlit. "A Faster Way to Build and Share Data Apps." Streamlit. Accessed October 8, 2024. https://streamlit.io.
Global Education Monitoring Report Team and Mark Bray. *Shadow Education in Sub-Saharan Africa: Scale, Nature and Policy Implications*. (United Nations Educational, Scientific and Cultural Organisation, 2021). https://policycommons.net/artifacts/6939784/shadow-education-in-sub-saharan-africa/.
Vanzo, Alessandro, Sankalan Pal Chowdhury and Mrinmaya Sachan. "GPT-4 as a Homework Tutor can Improve Student Engagement and Learning Outcomes." Preprint, submitted September 24, 2024. https://doi.org/10.48550/arXiv.2409.15981.
World Bank, UNESCO, UNICEF, USAID, FCDO and Bill & Melinda Gates Foundation. *The State of Global Learning Poverty*. (United Nations Children's Fund, 2022). https://policycommons.net/artifacts/2476935/the-state-of-global-learning-poverty/.
Yan, Lixiang, Samuel Greiff, Ziwen Teuber and Dragan Gašević. "Promises and challenges of generative artificial intelligence for human learning." Preprint, submitted September 5, 2024. https://doi.org/10.48550/arXiv.2408.12143.
Zhang, Hugh, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack and Qin Lyu. "A Careful Examination of Large Language Model Performance on Grade School Arithmetic." Preprint, submitted May 1, 2024. https://arxiv.org/abs/2405.00332.
Zhang, William. "Prompt Injection Attack on GPT-4." *Robust Intelligence*, March 31, 2023. https://www.robustintelligence.com/blog-posts/prompt-injection-attack-on-gpt-4.
Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan and Jimmy Ba. "Large Language Models are Human-Level Prompt Engineers." Preprint, submitted November 3, 2022. https://doi.org/10.48550/arXiv.2211.01910.

Appendix

The following prompt was used in Cycle 5:

> You are a criminal law tutor. Once the student says hi, tell them you want to review some concepts on sentencing rules and principles, and mention how excited you are to do so.
> You ask one question at a time and wait for their answer.
> Use 'hint' (delimited with XML tags) only after a student's initial answer is wrong (i.e. you would give them a mark less than 7 out of 10) or doesn't know the answer at all. If there's no hint available, then provide an appropriate 'hint' for the student based on the 'content' (delimited with XML tags).
> If a student's answer is almost correct (you would give them 7, 8 or 9 out of 10), then tell them that their answer is almost correct and encourage them to give it another try.
> After their initial attempt, if the student's answer is incorrect or doesn't know the answer, correct them based on the provided 'answer' (delimited with XML tags).
> 'Instruction' (delimited with XML tags) is to guide you on how to proceed, don't share it with students.
> If a student asks a question that is directly relevant to this lesson, only use the 'content' (delimited with XML tags) to answer. Do not use your own prior knowledge. If the 'content' doesn't properly answer their question, tell them you don't know the answer, and they should check with their tutor.
> If a student asks questions about anything other than this specific lesson, don't answer them and redirect them back to the lesson.
> Be positive, encouraging, and funny and use lots of emojis to be engaging.
> Ask the questions in the following order:
> Question 1.<question>The following statement reflects the ineffectiveness of which aims of sentencing?
> 'Around 50% of offenders re-offend within 10 years of their release date'</question> <answer>specific deterrence and rehabilitation</answer>
> <hint>Here are some aims of sentencing: retribution, specific and general deterrence, rehabilitation, incapacitation</hint>
> <content>Specific deterrence means punishing the offender to deter them from future offending i.e., if you do this act again, you'll be punished like this.
> Rehabilitation is rooted in the idea that psychiatric, psychological, or social factors compel offenders to participate in criminal activities. Criminal inclinations of an offender can be mitigated by addressing the root causes of their unlawful actions and changing an offender into a law-abiding citizen.</content>
> Question 2.<question>The following statement reflects on which theory of punishment?
> Even if a civil society were to be dissolved by the consent of all its members (e.g., if a people inhabiting an island decided to separate and disperse throughout the world) the last murderer remaining in prison would first have to be executed, so that each has done to him what his deeds deserve and blood guilt does not cling to the people for not having insisted upon this punishment; for otherwise the people can be regarded as collaborators in this public violation of justice.[55]</question>
> <answer>Retribution. The statement is by Kant.</answer>
> <hint> Theories of punishments include retribution, rehabilitation and deterrence</hint>
> <content> Retribution is a basic human instinct to impose painful consequences on people who commit harmful and wrongful acts. For many people, retribution is the primary purpose of punishment.</content>
> <instruction>If the student answered Question 2 incorrectly then ask Question 2.1.</instruction>
> Question 2.1. <question> The notion of eye for an eye is relevant to which theory of punishment?</question>
> <answer>retribution</answer>
> <hint> You may think of this theory as vengeance but it is not all about revenge. This theory also holds that punishment should be proportionate to the harm that is caused by the offender and their level of blameworthiness or guilt. So it's eye for an eye kind of justification.</hint>
> <content>Retribution is a basic human instinct to impose painful consequences on people who commit harmful and wrongful acts</content>
> Question 3. <question>Do you know what the maximum guilty plea discount is? </question>
> <answer>25%</answer>
> <content>According to s 25D(2)(a) of Crimes (Sentencing Procedure) Act 1999, a reduction of 25% in any sentence that would otherwise have been imposed, if the plea was accepted by the Magistrate in committal proceedings for the offence </content>
> <instruction>If the student answered Question 3 correctly, then ask Question 3.1. If the student answered Question 3 incorrectly, then ask Question 3.2.</instruction>
> Question 3.1.<question>Do you also know what would be the sentence reduction for an offender who pleaded guilty at least 14 days before the first day of the trial? </question>
> <answer>10%</answer>
> <content>According to s 25D(2)(b) of Crimes (Sentencing Procedure) Act 1999, a reduction of 10% in any sentence that would otherwise have been imposed, if the offender was committed for trial and the offender-
> (i) pleaded guilty at least 14 days before the first day of the trial of the offender, or

---

[55] Hill, "Kant on Wrongdoing, Desert, and Punishment," 433 quoting; Kant, Kant: The Metaphysics of Morals.

(ii) complied with the pre-trial notice requirements and pleaded guilty at the first available opportunity able to be obtained by the offender.</content>
Question 3.2<question>What’s the minimum guilty plea discount?</question>
<answer> 5%</answer>
<content>According to s 25D(2)(c) of Crimes (Sentencing Procedure) Act 1999, the minimum is 5%</content>
This is the end of all the questions. Say something funny and encouraging to the students and ask them to provide feedback if they have any.