Bond University

Legal Education Review

Volume 33 Issue 1

2023

‘Words are Flowing Out Like Endless Rain Into a Paper Cup’: ChatGPT & Law School Assessments

Stuart Hargreaves
Chinese University of Hong Kong

‘WORDS ARE FLOWING OUT LIKE ENDLESS RAIN INTO A PAPER CUP’:* CHATGPT & LAW SCHOOL ASSESSMENTS

STUART HARGREAVES+

I INTRODUCTION

ChatGPT is one of a number of ‘generative AI’[1] tools that appeared in 2022 (others have included tools that would help users create images and digital art,[2] video[3], generative music,[4] creative storytelling,[5] avatars delivering corporate training materials,[6] and more).[7] ChatGPT is able to generate written English text of a remarkably high quality; in one study, human readers found it difficult to distinguish a 500 word news article written by the underlying AI that powered it from one written by a human.[8] One philosophy professor specializing in AI

---

* Lennon/McCartney, ‘Across the Universe’, Apple Records 1969.

+ Faculty of Law, Chinese University of Hong Kong. My thanks to the many colleagues who graciously agreed to give their time and expertise to participate in this project: Rehan Abeyratne; Anatole Boute; Kevin Cheng; Matt Cheung; Agnes Chong; Bryan Druzin; Elliot Fung; Steve Gallagher; Stephen Hall; Dennis Hie; Queenie Lai; Jae Woon Lee; Michael Lower; Luke Marsh; Michelle Miao; Michael Ramsden; Peter Rhodes; Paul Schmidt; Samuli Seppanen; Jingyi Wang; Lutz-Christian Wolff.

1 This refers to an artificial intelligence model able to generate new and appropriate content of varying forms in response to simple language-based inputs from a user.

2 See eg *Stable Diffusion Online* (Web Page) <https://stablediffusionweb.com>; ‘DALL-E 2’, *OpenAI* (Web Page) <https://openai.com/dall-e-2/>; ‘Text to Image – AI Image Generator API’, *DeepAI* (Web Page) <https://deepai.org/machine-learning-model/text2img>; ‘AI Art Generator’, *Hotpot* (Web Page) <https://hotpot.ai/art-generator>; ‘DreamFusion: Text-to-3D Using 2D Diffusion’, *DreamFusion* (Web Page) <https://dreamfusion3d.github.io>.

3 ‘Imagen Video’, *Google* (Web Page) <https://imagen.research.google/video/>; ‘Create AI Videos from Text’, *ELAI* (Web Page) <https://elai.io>.

4 ‘Jukebox’, *OpenAI* (Web Page) <https://openai.com/blog/jukebox/>; ‘MusicLM: Generating Music From Text’, *Google* (Web Page) <https://google-research.github.io/seanet/musiclm/examples/>.

5 *Wordcraft Writers Workshop* (Web Page) <https://wordcraft-writers-workshop.appspot.com>.

6 ‘Create Videos from Plain Text in Minutes’, *Synthesia* <https://synthesia.io>

7 Haomaio Huang, ‘The Generative AI Revolution Has Begun – How Did We Get Here?’ *ArsTechnica* (Web Page, 30 Jan. 2023) <https://arstechnica.com/gadgets/2023/01/the-generative-ai-revolution-has-begun-how-did-we-get-here/>.

8 Tom Brown et al., ‘Language Models are Few Shot Learners’ (2020) *Computation & Language arXiv:2005.14165*, 16.

declared that its output was so good as to be 'disconcerting', and closer to passing the Turing test[9] than any previous system.[10] From an end-user's perspective, one of the most remarkable aspects is the speed with which ChatGPT answers a question put to it. Unless the server is under heavy load the response to a query – no matter how long or seemingly complicated – is nearly instant. Once a prompt[11] is received, the 'answers' simply begin to flow down the screen – inspiring the title of this paper. If one anthropomorphized ChatGPT, one would say it appears very confident: it instantly gives an answer of an appropriate length, written in accurate prose. There is no hesitation, no pause to gather thoughts or contemplate. It is a highly seductive process; it is easy to imagine the system 'knows' exactly what it is talking about – and yet, as this paper shows, it often does not.

The remarkable ease with which ChatGPT can generate natural sounding text across a range of domains in response to natural language prompts made it an instant online sensation, with over 100 million unique users and 590 million visits within less than two months of launch.[12] AI-generated text that was often indistinguishable from the real thing promised to assist with an array of tasks from the creative to the mundane.[13] End-users adopted ChatGPT to complete job application tests,[14] come up with Christmas list ideas,[15] make workout plans,[16] debug computer code,[17] engage in conversations potential romantic partners online,[18] lower bills by negotiating with customer

---

9 Alan Turing's 'imitation game' proposed that even if machines cannot 'think', an important test of equivalence to human intelligence would be if humans could interact with a machine and not realize they were not interacting with another person. See 'The Turing Test', *Stanford Encyclopedia of Philosophy* (Web Page, 4 October 2021) <https://plato.stanford.edu/entries/turing-test/>.

10 David Chalmers, 'GPT-3 and General Intelligence', *Daily Nous* (Blogpost, 30 July 2020) <https://dailynous.com/2020/07/30/philosophers-gpt-3/#chalmers>.

11 A 'prompt' refers to whatever input the user enters into the system; they may be very short or very detailed, and have a well-crafted prompt appears an important aspect of getting the desired for response.

12 Dan Milmo, 'ChatGPT Reaches 100 million Users Two Months After Launch', *The Guardian* (online, 2 February 2023) <https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app>.

13 Monica White, 'Top 10 Most Insane Things ChatGPT Has Done This Week', *Springboard* (online, 9 December 2022) <https://www.springboard.com/blog/news/chatgpt-revolution/>.

14 Tom Acres, 'Recruitment Team Unwittingly Recommends ChatGPT for a Job Interview', *Sky News* (online, 17 January 2023) <https://news.sky.com/story/recruitment-team-unwittingly-recommends-chatgpt-for-job-interview-12788770>.

15 Alan Truly, 'I Used the ChatGPT AI Chatbot to Do My Holiday Shopping this Year', *Digital Trends* (online, 13 December 2022) <https://www.digitaltrends.com/computing/i-used-chatgpt-to-do-my-holiday-shopping-this-year/>.

16 @anothercohen (Twitter, 5 December 2022, 6:28 AM AWST) <https://twitter.com/anothercohen/status/1599531037570502656>.

17 @jdkelly (Twitter, 1 December 2022, 2:29 AM AWST) <https://twitter.com/jdjkelly/status/1598021488795586561>.

18 Anna Iovine, 'Tinder Users Are Using ChatGPT to Message Matches', *Mashable* (online, 17 December 2022) <https://mashable.com/article/chatgpt-tinder-tiktok>.

service,[19] prepare interview questions,[20] negotiate with insurance providers,[21] write poetry,[22] and generate original creative works in the style of others.[23] One programmer used ChatGPT to answer a number of publicly available exams in the United States: the version of the system used in this project answered 70% of the questions on the Medical Licensing Examination correctly, answered 35/50 correctly on a sample section of the Multistate Bar Exam, 9/15 on a sample of the Multistate Professional Responsibility Exam, scored 78% on the multiple-choice portion of a high school chemistry exam, and scored 149 (40th percentile) on a previous version of the Law School Admission Test.[24] ChatGPT can be put to harmful uses too. It has already been used to create malware,[25] and is clear potential for the generation of misinformation at scale, since ChatGPT creates superficially believable responses to inputs. Indeed, a coding question and answer website banned AI-generated content after it was overwhelmed with 'false' answers.[26]

While others have also considered the effect of generative AI on assessments in higher education,[27] to the author's knowledge this paper is the first to consider ChatGPT's use to complete a wide range of common law, English-language law school exams.[28] In Part 2, I explain the basics of ChatGPT. In Part 3, I situate ChatGPT in the context of

---

19 Emma Roth, 'DoNotPay Is Launching an AI Chatbot That Can Negotiate Your Bills', *The Verge* (online, 13 December 2022) <https://www.theverge.com/2022/12/13/23505873/donotpay-negotiate-bills-ai-chatbot>.

20 @sethbannon (Twitter, 1 December 2022, 3:28 AM AWST) <https://twitter.com/sethbannon/status/1598036175285276672>.

21 @stuartblitz (Twitter, 14 December 2022, 9:13 AM AWST) <https://twitter.com/StuartBlitz/status/1602834224284897282>.

22 Jack Cushman, 'ChatGPT: Poems and Secrets', *Library Innovation Lab* (Web Page, 20 December 2022) <https://lil.law.harvard.edu/blog/2022/12/20/chatgpt-poems-and-secrets/>.

23 For example, one user asked ChatGPT to write a new song in the style of Nick Cave: 'Issue #218', *The Red Hand Files* (Web Page, January 2023) <https://www.theredhandfiles.com/chat-gpt-what-do-you-think/>. Nick Cave, however, did not approve of the results: Sian Cain, ''This Song Sucks': Nick Cave Responds to a ChatGPT Song Written in the Style of Nick Cave', *The Guardian* (online, 17 January 2023) <https://www.theguardian.com/music/2023/jan/17/this-song-sucks-nick-cave-responds-to-chatgpt-song-written-in-style-of-nick-cave>.

24 @pythonprimes (Twitter, 11 December 2022, 3:46 AM AWST) <https://twitter.com/pythonprimes/status/1601664776194912256>.

25 Alessandro Mascellino, 'ChatGPT Creates Polymorphic Malware', *InfoSecurity* (online, 18 January 2023) <https://www.infosecurity-magazine.com/news/chatgpt-creates-polymorphic-malware/>.

26 'Temporary Policy: ChatGPT Is Banned', *Stack Overflow* (Web Page, 5 December 2022) <https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned>.

27 See eg Xiaoming Zhai 'ChatGPT User Experience: Implications for Education' (27 December 2022) <https://ssrn.com/abstract=4312418>; Margaret Ryznar, 'Exams in the Time of ChatGPT' (27 January 2023) <https://ssrn.com/abstract=3684958>. One also focused on performance in law exams: Jonathan Choi et al, ChatGPT Goes to Law School' (23 January 2023) <https://ssrn.com/abstract=4335905>.

28 Choi et al (n 27) considered only four courses. It also did so in the context of a US law school which, as will I argue in Part 5, may result in different outcomes due to an Anglo-American bias in the training data upon which ChatGPT is trained.

existing literature regarding academic integrity. In Part 4, I describe the methodology and results of my research and note that training data is a particular challenge for law assessments in smaller jurisdictions. In Part 5, I consider the implications of ChatGPT for the future of teaching and learning in higher education generally. Appendix I includes summaries of all 24 exams put through the system, the grades they were assigned by the expert assessor, and general feedback they provided about ChatGPT's performance. While the technology is incredibly impressive, this research shows that ChatGPT often gives entirely *incorrect* answers in the legal context. It invents cases. It fails to spot obvious issues. It ignores applicable law. The idea that a judge could rely on ChatGPT to summarize an area of law for use in the court remains, for the time being, a poor idea.[29]

II THE BASICS OF GENERATIVE AI & CHATGPT

OpenAI was founded in 2015 as a non-profit with a stated plan to 'advance digital intelligence in a way that is likely to benefit humanity as a whole, unconstrained by a need to generate financial return.'[30] Funders committed an initial USD $1 billion before a single product had been released. In 2019, OpenAI transitioned to a 'capped' for-profit model that would allow it to more easily raise capital, apparently concluding that the initial seed money would be quickly exhausted.[31] Technology giant Microsoft announced a USD $10 billion investment and deep partnership with OpenAI, following on from two earlier $1 billion investments in 2019 and 2021.[32] While the research preview was free, OpenAI's head indicated that it would eventually have to be monetized as the 'compute costs [were] eye-watering.'[33] In February 2023, OpenAI announced the launch of ChatGPT Plus, a USD $20/month premium subscription service that would guarantee access to the service even at peak times, and promised faster responses and

---

29 Cf Luke Taylor, 'Colombia Judge Says He Used ChatGPT in Ruling', *The Guardian* (online, 3 February 2023) <https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling>.

30 'Introducing OpenAI', *OpenAI* (Web Page, 11 December 2015) <https://openai.com/blog/introducing-openai/>.

31 'OpenAI LP', *Open AI* (Blogpost, 11 March 2019) <https://openai.com/blog/openai-lp/>; Devin Coldewey, 'OpenAI Shifts From Nonprofit to 'Capped-Profit' to Attract Capital', *TechCrunch* (online, 12 March 2019) <https://techcrunch.com/2019/03/11/openai-shifts-from-nonprofit-to-capped-profit-to-attract-capital/>.

32 Dina Bass, 'Microsoft Invests $10 Billion in ChatGPT Maker OpenAI', *Bloomberg* (online, 23 January 2023) <https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai>.

33 @sama (Twitter, 5 December 2022 3:35PM AWST) <https://twitter.com/sama/status/1599668808285028353>.

early access to new features.[34] OpenAI estimated its revenue to be $1b annually by 2024.[35]

OpenAI released its first 'Generative Pre-trained Transformer' (GPT) in 2016. Earlier attempts at natural language processing depended upon manually-labelled data – in other words, humans had to teach the system what something meant. In contrast, generative pre-training is a method that allows for significant improvement in the ability of natural language processing models to 'learn effectively from raw text… alleviating the dependence on supervised learning.'[36] This breakthrough meant natural language processing tools could be trained on much larger amounts of entirely unstructured data.[37] GPT1 was based on 117 million parameters (the weights and layers of the underlying neural network that is used for training the model),[38] trained on around 5 gigabytes (GB) of text.[39]

Another version of GPT (GPT2) was released that year, and consisted of 1.5 billion parameters,[40] trained on 40GB of text.[41] The third version of this tool (GPT3, released in 2020) was larger by an order of magnitude: 175 billion parameters, trained upon roughly 45 *tera*bytes of text.[42] This was composed of four sources: Common Crawl (filtered) (60% of total material), WebText2 (22%), Books1 (8%), Books2 (8%), and Wikipedia (3%).[43] Further explanation of these sources is helpful as in a sense they ultimately dictate what ChatGPT 'knows' and shape its outputs. Common Crawl is a non-profit organization that uses web crawling tools to essentially archive as much of the open internet as it can.[44] It is open-source and contains 12 years of data, including 'raw web page data, metadata extracts, and text extracts' totalling in the petabytes.[45] OpenAI 'filtered' this data through the application of machine learning tools to identify 'high quality' documents and remove duplicates and overlap with the other sources.[46]

---

34 'Introducing ChatGPT Plus', *OpenAI* (Blogpost, 1 February 2023) <https://openai.com/blog/chatgpt-plus/>.

35 Jeffrey Dastin, Krystal Hu, and Paresh Dave, 'Exclusive: ChatGPT Owner OpenAI Projects $1 Billion in Revenue by 2024', *Reuters* (online, 15 December 2022) <https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/>.

36 Alec Radford et al., 'Improving Language Understanding by Generative Pre-Training', *OpenAI* (Internal Research Paper, 11 June 2018) <https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf>.

37 Huang (n 8).

38 Priya Shee, 'The Journey of Open AI GPT Models', *Medium* (online, 10 November 2020) <https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2>.

39 'Datasets: Bookcorpus', *Hugging Face* (Digital Repository) <https://huggingface.co/datasets/bookcorpus>.

40 Huang (n 8).

41 Shee (n 37).

42 Brown et al. (n 9).

43 Ibid 8.

44 'Frequently Asked Questions', *Common Crawl* (Webpage) <https://commoncrawl.org/big-picture/frequently-asked-questions/>.

45 'Want To Use Our Data?', *Common Crawl* (Webpage) <https://commoncrawl.org/the-data/>.

46 Brown et al. (n 9) 43.

WebText2 is an 'expanded version'[47] of the original WebText, an OpenAI-developed corpus initially designed for GPT2 that was based upon scraping the content of webpages linked to by Reddit posts that had received more than three upvotes (in attempt to ensure only 'quality' pages were included).[48]

Researchers have found content of the Books1 and Books2 (like WebText, both generated internally by OpenAI) hard to determine, but it is not a database of all public domain books.[49] Presser has suggested that Books1 is similar to the bookcorpus dataset,[50] an open-source collection of around 7000 of novels primarily but not exclusively free from copyright by largely unpublished authors, with a strong bias towards the genre of fictional romance.[51] OpenAI has declined to reveal the contents of Books2, though Presser speculates it might be similar to 'libgen', an unauthorized collection of over two million copyrighted materials including textbooks and scholarly articles.[52] Wikipedia is the well-known online 'knowledge repository' entirely dependent upon the contributions of its users; at the time of writing it consists of nearly 7 million English-language content pages.[53]

These datasets do not contribute equally to GPT3. According to OpenAI, the filtered Common Crawl dataset accounts is given 60% weight in the training mix, WebText2 for 22%, Books1 and Books 2 each for 8%, and Wikipedia for 3%.[54] Weight in the training mix 'refers to the fraction of examples during training from a given dataset, which... is not proportional to the size of the dataset.'[55] Roberts' analysis demonstrates that because of the wide variance on the size of the different datasets, this choice essentially amplifies the importance of WebText2 and Wikipedia in training.[56] This matters, since 'the choice

---

47 Ibid 8.

48 Alec Radford et al, 'Language Models Are Unsupervised Multitask Learners', *OpenAI* (Internal research paper, 27 May 2020) <https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf> (unpaginated).

49 Greg Roberts, 'AI Training Datasets: TEXT Contents', *Musings of Freedom* (Blogpost) <https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/#AI_Training_Datasets_TEXT_contents>; @theshawwn (Twitter, 25 October 2020 4:32 PM AWST) <https://twitter.com/theshawwn/status/1320282151645073408>.

50 @theshawwn (Twitter, 25 October 2020 4:32 PM AWST) <https://twitter.com/theshawwn/status/1320282151645073408>.

51 Jack Bandy, 'Dirty Secrets of BookCorpus, a Key Dataset in Machine Learning', *Medium* (Blogpost, 12 May 2021) <https://towardsdatascience.com/dirty-secrets-of-bookcorpus-a-key-dataset-in-machine-learning-6ee2927e8650>.

52 @theshawwn (Twitter, 25 October 2020 4:32 PM AWST) <https://twitter.com/theshawwn/status/1320282151645073408>; Balazs Bodo, Daniel Antal, and Zoltan Puha, 'Can Scholarly Pirate Libraries Bridge the Knowledge Access Gap? An Empirical Study on the Structural Conditions of Book Piracy in Global and European Academia' (2020) 15(2) *PLoS One* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7714232/>.

53 'Statistics', *Wikipedia* (Webpage) <https://en.wikipedia.org/wiki/Special:Statistics>.

54 Brown et al. (n 9) 9.

55 Ibid.

56 Roberts (n 69).

and careful selection of raw intake knowledge feeds could drastically affect the resultant AI 'brain' that develops.'[57]

This means that flaws or inadequacies in AI training data lead to the risk of biased answers, a separate issue from correctness. This is a well-studied phenomena.[58] In short, since the data that comprises the training mix is drawn from the open web, unless great care is taken then the answers a large language model generates will reflect the biases contained within that data. OpenAI acknowledges that ChatGPT can produce 'harmful instructions or biased content', but the only solution it provides is that users may click a 'thumbs-down' button if they think the answer is incorrect. It is not clear to what extent (or how) OpenAI will take into account this kind of user feedback on issues upon which its users will necessarily have diverse views.

In any event, the massive increase in training data between GPT2 and GPT3 meant the latter was far superior at a range of natural language processing tasks and generating outputs that appeared 'human-like'.[59]

> OpenAI researchers discovered that in making the models bigger, they didn't just get better at producing text. The models could learn entirely new behaviors simply by being shown new training data. In particular, the researchers discovered that GPT3 could be trained to follow instructions in plain English without having to explicitly design the model that way… So instead of creating single-purpose language tools, GPT3 is a multi-purpose language tool that can be easily used in many ways by many people without requiring them to learn programming languages or other computer tools.[60]

OpenAI took GPT3 and fine-tuned it through human evaluations of competing outputs; those evaluations were then incorporated into a reward training model, significantly improving the performance of

---

57 Ibid.

58 See eg James Zou and Londa Schiebinger, 'AI Can Be Sexist and Racist—It's Time to Make It Fair' (2018) 559(7714) *Nature* 324; Natalia Norori et al., 'Addressing Bias in Big Data and AI For Health Care: A Call For Open Science' (2021) 2(10) *Patterns* 100347; Drew Roselli, Jeanna Matthews, and Nisha Talagala, 'Managing Bias in AI' in *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)* (Association for Computing Machinery, 2019) 539–544; Susan Leavy, Barry O'Sullivan, and Eugenia Siapera, 'Data, Power and Bias in Artificial Intelligence', *Computer Science: Computers & Society* (online, 28 July 2020) <https://doi.org/10.48550/arXiv.2008.07341>; Eirini Ntoutsi et al., 'Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey', *WIREs Data Mining Knowledge Discovery* (online, 3 February 2020) <https://doi.org/10.1002/widm.1356>; Ramya Srinivasan and Ajay Chander, 'Biases in AI Systems' (2021) 64(8) *Communications of the ACM* 44; Jake Silberg and James Manyika, 'Notes From the AI Frontier: Tackling Bias in AI (and in Humans)', *McKinsey Global Institute* (online, June 2019) <https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/MGI-Tackling-bias-in-AI-June-2019.pdf>.

59 Kindra Cooper, 'OpenAI GPT-3: Everything You Need to Know', *Springboard* (online, 1 Nov. 2021) <https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/>.

60 Huang (n 8).

GPT3 in creating outputs that 'aligned with human intent.'[61] OpenAI called this fine-tuned version of GPT3 'InstructGPT' or GPT3.5. ChatGPT was released as 'research preview' in November 2022, designed to allow end-users to interact with the underlying GPT3.5/InstructGPT model in a conversational way.

> The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.[62]

The technology is rapidly evolving, and openAI's rivals are not standing still. Between the time this project was conducted (and the bulk of the paper was written) and the time it was published, Microsoft incorporated ChatGPT directly into its Bing search engine,[63] Google launched a competitor called Bard[64] after deeming ChatGPT a 'code red' threat to its primary business model, [65] Baidu announced a competitor called Ernie (though did not make it widely available to the public), [66] and openAI itself launched a new version (GPT4) that allowed for multi-modal inputs [67] and promised greater accuracy in results.[68] At launch, openAI reported this new version scored in the 90th percentile on the Uniform Bar Exam (vs 10th percentile for GPT3.5), 163 on the LSAT (vs 149), the 99th percentile on the verbal portion of the GRE (vs 63rd), and so on.[69]

There is now a highly competitive, rapidly evolving landscape of large language models, and companies seem to be releasing flawed projects just to stay in the conversation. Google was criticized by its own employees for releasing Bard 'too early' just to try and compete with openAI. [70] Likewise Baidu's shares dropped by 10% after an unimpressive demo of Ernie was apparently rushed out to prove it could

---

61 Long Ouyang et al., 'Training Language Models to Follow Instructions with Human Feedback', *OpenAI* (Internal Research Paper, 4 March 2022) <https://arxiv.org/pdf/2203.02155.pdf>.

62 'ChatGPT: Optimizing Language Models for Dialogue', *OpenAI* (Blogpost, 30 November 2022) <https://openai.com/blog/chatgpt/>.

63 'Microsoft Unveils New Bing with ChatGPT Powers', *BBC News* (online, February 7, 2023) < https://www.bbc.com/news/business-64562672>.

64 'Google Launches Bard AI Chatbot to Counter ChatGPT', *Wall Street Journal* (online, March 21 2023) < https://www.wsj.com/articles/google-launches-bard-ai-chatbot-to-counter-chatgpt-2200c357>.

65 Nico Grant and Cade Metz, 'A New Chat Bot is a 'Code Red' for Google's Search Business', *The New York Times* (online, 21 December 2022) <https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html>.

66 Lyric Li and Meaghan Tobin, 'Ernie Bot, China's Answer to ChatGPT, Is Delayed – Again', *The Washington Post* (online, March 28 2003) <https://www.washingtonpost.com/world/2023/03/28/china-baidu-chatbot-ai-ernie/>.

67 In other words, GPT4 can understand not just text inputs, but pictures as well.

68 'GPT-4', *OpenAI* (Blogpost, March 14 2023) <https://openai.com/research/gpt-4>.

69 Ibid.

70 Cecily Mauran, 'Google Employees Also Think the Bard Launch Was 'Botched' and Rushed', *Mashable* (online, February 11 2023) <https://mashable.com/article/google-employees-memegen-internal-forum-criticize-bard-launch>.

compete in this space.[71] Competition also seems to have led to decreased transparency. While openAI had identified details about the training data and other basic facts of earlier models, upon the release of GPT4, it declared it would no longer do so:

> Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.[72]

Though some have declared these systems may have a social and economic impact on par with the development of the Gutenberg Press or possibly even 'the wheel',[73] the regulatory landscape has struggled to keep pace with these rapid changes. In March 2023, an open letter penned by tech leaders and AI experts called for a six-month pause on all further developments of systems like ChatGPT until appropriate regulatory guardrails were in place.[74] OpenAI itself seems to acknowledge that ChatGPT's rapid evolution suggests they are on the path towards successfully developing 'artificial general intelligence' (AGI)[75], and that may bring 'massive risks' (though at the time of writing they plan to continue nonetheless – just with caution).[76] Despite these concerns, the sole government intervention with ChatGPT at the time of writing has been over privacy concerns – the Italian data protection authority argued that openAI's approach to using scraped webdata for training without the consent of those who had generated that data was a violation of European data privacy law and ordered OpenAI to stop offering the service in Italy.[77]

Without question, ChatGPT and other forms of generative AI raise a range of important issues worthy of future research – potential bias in inputs and consequently outputs, the potential for widespread elimination of a range of knowledge work jobs, possible violations of privacy and copyright law as a result of the way training data is gathered, and the immense societal changes that would be wrought by the successful development of AGI. But this paper focuses on a discrete

---

71 Ryan McMorrow and Qianer Liu, 'Baidu Shares Fall After Ernie AI Chatbot Demo Disappoints', *ArsTechnica* (online, March 16 2023) <https://arstechnica.com/information-technology/2023/03/chinese-search-giant-launches-ai-chatbot-with-prerecorded-demo/>.

72 'GPT-4 Technical Report', *OpenAI* (online, March 27 2023) <https://cdn.openai.com/papers/gpt-4.pdf>.

73 @business (Twitter, 12 December 2022 9:52AM AWST) <https://twitter.com/business/status/1602119041987977217>.

74 'Pause Giant AI Experiments: an Open Letter', *Future of Life Institute* (Blogpost) <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

75 'An Executive Primer on Artificial General Intelligence', *McKinsey & Company* (online, April 29 2020) <https://www.mckinsey.com/capabilities/operations/our-insights/an-executive-primer-on-artificial-general-intelligence>.

76 'Planning for AGI and Beyond', *OpenAI* (online, February 24 2023) <https://openai.com/blog/planning-for-agi-and-beyond>.

77 Elvira Pollina and Supantha Mukherjee, 'Italy Curbs ChatGPT, Starts Probe Over Privacy Concerns', *Reuters* (online, April 1 2023) <https://www.reuters.com/technology/italy-data-protection-agency-opens-chatgpt-probe-privacy-concerns-2023-03-31/>.

issue – the use (or mis-use) of generative AI to answer exam questions in higher education settings.

III USING CHATGPT TO ANSWER LAW EXAMS

A *Background and Methodology*

The Faculty of Law of the Chinese University of Hong Kong ('CUHK Law') ranked as one of the top 50 law schools in the world in the 2023 Times Higher Education rankings.[78] It is one of three law schools in Hong Kong, a common law jurisdiction under Chinese sovereignty. CUHK Law offers a number of taught legal degrees including the LLB and the JD (both entry-level law degrees that serve as the first step towards professional qualification), the PCLL (a skills-based year-long programme that must also be completed by all aspiring barristers and solicitors in Hong Kong), and the LLM (which does not lead to professional qualification). The medium of instruction for all degrees is English.

At the outset, it should be noted that ChatGPT is not formally available in Hong Kong at the time of writing. Registration requires use of a VPN to spoof user location, as well as access to an overseas phone number with an area code matching that location. While these hurdles are an annoyance, it is reasonable to assume that a) an interested student would be able to surmount these hurdles; b) the service will in any event expand to more locations over time as OpenAI's resources increase; and c) competitors will emerge who will offer similar and perhaps even superior services globally. In any event, the lessons drawn from this project are applicable across many jurisdictions where ChatGPT and other forms of generative AI are readily available.

This project solicited past exams (or portions thereof) from academic staff delivering courses across the taught programmes. The plan was simple – the project would put those exam questions through ChatGPT and send the results back to the academic staff concerned for evaluation. The project design and methodology has both strengths and weaknesses. On the one hand, by soliciting questions from various courses across four different programmes of study, it ensured there were a variety of assessment models that could be tested – essays, problem-based issue spotting questions, true/false, and so on. Reliance on past exam questions meant also that the project could begin immediately rather than having to wait for the next exam period. On the other hand, since participation was voluntary, only a fraction of the exams assessed annually across CUHK Law were put through the system and there was no way to ensure an even spread between the various programmes.

Moreover, the project only measures the exam-answering capability of ChatGPT at a singular point in time – all exams were answered by

---

[78] 'World University Rankings 2023 by Subject: Law', *Times Higher Education* (Webpage) <https://www.timeshighereducation.com/world-university-rankings/2023/subject-ranking/law#!>.

the GPT3.5 model, specifically the 'January 9, 2023' version.[79] As noted above,[80] GPT4 was released shortly after this paper was first submitted for consideration. A further methodological challenge can be found in the style of prompts used. Since the project is aimed at understanding how a student might use ChatGPT to answer an exam question, the author had to consider how to act out this role. Because ChatGPT is a conversation-based tool it was not always possible to simply cut and paste large exams into the input box in a single go. Deciding how to the 'share' the information with ChatGPT was an open question. Should it be divided up into chunks? Should the system be primed by identifying the topic? In attempt to maintain consistency, the project proceeded by starting a fresh question with ChatGPT for each exam, and primed the system by saying something to the effect of 'I will give you a hypothetical scenario and then will ask you questions about it that relate to [Hong Kong contract law].' While on one occasion[81] a question was rephrased when ChatGPT had seemingly completely misunderstood it, in general the output was sent to the assessor even if contained an obviously suspect legal analysis. What the project offers, then, is a snapshot of 'bare-minimum' results achieved by using ChatGPT to answer law exams in early 2023. Over time, not only will the system (or its competitors) improve, but students will likely become more proficient at crafting prompts that generate more detailed answers.[82]

A separate methodological concern is that the project was not 'blind'. Without question, it would have been superior to have assigned ChatGPT a fake student number and a complete range of exams to be marked during the actual exam period by assessors who thought they were evaluating 'just another student'. However, such a project would have taken not only much more time (at the time of writing, the next exam period will be in May 2023), but also significantly more institutional buy-in. So, the assessors of the exams were well aware that they were being asked to evaluate answers written by an AI. That may have led some to skim the answers quickly; after all, assessing exams is time-consuming and not particularly enjoyable, and I was asking them to do more work for my own benefit. It also may have led to some unconscious bias against the answers – a surprising number of colleagues who agreed to participate in the project pre-emptively voiced sentiments along the lines of 'I doubt the system will be able to handle *my* exam', 'surely it will not be able to answer problem questions', or 'I hope it fails', and the like. There appeared to be a belief amongst

---

79 The system is regularly updated to fix bugs, errors, and take into account received feedback. There is no way for an end-user to roll-back to a previous version: 'ChatGPT – Release Notes', *OpenAI* (Blogpost, 13 February 2023) <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>.

80 See above n 67.

81 Criminal Law, LLB. See Appendix I for futher explanation.

82 Indeed, jobs for remarkably well-compensated 'prompt engineers' have already sprung up. See eg Conrad Quilty-Harper, '$335,000 Pay for 'AI Whisperer' Jobs Appears in Red-Hot Market', *Bloomberg* (online, March 29 2023) <https://www.bloomberg.com/news/articles/2023-03-29/ai-chatgpt-related-prompt-engineer-jobs-pay-up-to-335-000#xj4y7vzkg>.

some that what was being evaluated was the *difficulty* of the exam from the perspective of an AI as some kind of proxy for *quality* – which of course is not the case.

What the project sought to evaluate was not whether an exam was 'hard', but the extent to which ChatGPT was able to answer varied forms of law assessment. Computers are good at certain tasks but not others. Software has been superior to the world's very best chess players since the early 1990s,[83] but designing a bipedal robot that can walk up and down a staircase it has never previously seen remains an incredible challenge.[84] Conversely, the average four year old human can easily perform the latter, but it would be incredibly rare to see a child playing chess at anything beyond a rudimentary level. What is easy for an AI may be a challenge for a human student, and vice-versa; this project does not (and indeed cannot) use ChatGPT to measure the objective difficulty of an exam, let alone the academic *bona fides* of the human who set it.

But despite its limitations, the results of the project are both valuable and interesting. They indicate that while certain forms of assessment may need to be re-thought, on balance ChatGPT significantly struggles with common assessment styles used in law schools. However, given systems like ChatGPT are certain to improve in the coming years, the reality is that the use of these tools is likely going force a rethink in not simply assessment, but teaching and learning in general; I consider this point in Part 5.

B *Results*

A complete summary of the exams run through the system and assessors' comments is found in Appendix I;[85] here I provide a high-level overview of the results. The exams can be grouped into four categories. Those with answers that were deemed to reach (at least in part) into the A range ('strong answers'), those that were primarily in the B range ('reasonably good answers'), those that received a C or D ('passable but poor answers'), and those that received an F ('failing answers'). It is worth remembering that if these were in-class exams, students would have three hours to write them. Take-home exams typically must be returned within 48 to 72 hours after release. This project took no more than 10 minutes to put a single exam through the system, craft revised prompts on occasion, and then cut and paste the results into a document to email back to the assessor. It is reasonable to assume that the average student could improve upon these results without too much effort.

---

[83] Larry Greenmeier, '20 Years After Deep Blue: How AI Has Advanced Since Conquering Chess', *Scientific American* (online, 2 June 2017) <https://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/>.

[84] Jonah Siekmann et al., 'Blind Bipedal Stair Traversal via Sim-to-Real Reinforcement Learning', *Computer Science: Robotics* (Research paper, 18 May 2021) <https://arxiv.org/abs/2105.08328>.

[85] Infra, 24.

Based on the judgment of expert assessors, ChatGPT performed best on Jurisprudence (JD),[86] International Tax Law, and International Environmental Law. These were the only exams deemed to possible reach into the 'A' range. What similarities do they share? First, they were all exclusively essay-based assessments, with questions that encouraged students to reflect upon or analyze relatively broad topics. The topics discussed were not connected to hyper-local intricacies of Hong Kong law. None of them referred obliquely to some technique or concept that could only be known if a student had attended a particular class. None invented scenarios to be considered. This then is the kind of assessment ChatGPT has the most success with, insofar as it can mimic a relatively well-written summation of ideas and concepts contained within its training data. It can offer critiques and contrasts as well, so long as those critiques can also be found within its training data. When topics are 'real' and 'international', there is a significantly higher likelihood that its training data will incorporate various perspectives on them.

The second category of exams that led to 'reasonably good answers' included Jurisprudence (LLB), Hong Kong Legal System, Legal System of the PRC, and Competition Law. In this band we can see that grades are lower when some of the factors considered above point in different directions.[87] In the case of the Hong Kong Legal System and Legal System of the PRC exams, both were exclusively essay-based. However, both required knowledge of legal systems that are less likely to be represented in ChatGPT's training data. Both failed to show sufficient detail in the answers according to the respective assessors, instead focusing largely on generalities about the respective legal systems.

However, since neither exam created hypothetical scenarios requiring the application of local laws to invented facts, ChatGPT was still able to obtain a B grade. Competition Law consisted of both an issue-spotting hypothetical pattern and two essay questions, with the former being notably weaker. ChatGPT was unable to correctly explain the relevant portion of the local Ordinance and made up a non-existent case in support of its answer. The essay questions referred to broader concepts in competition law including how those concepts were implemented in non-local instruments, which again presumably meant ChatGPT had more relevant material in its training database to draw upon.

The third category of 'poor but passable' answers included those given in International Business Transactions, Commercial Practice,

---

[86] The majority of courses run in both the LLB and JD programmes. Where exams from the same course across the two programmes were analyzed, this distinction is noted just for clarity. The same approach is taken where two exams from the same course across two different years were submitted for consideration.

[87] Jurisprudence (LLB) is an exception to this pattern. Though stylistically similar to exams in the first category, it nonetheless received a lower grade than its JD counterpart. While this was in large part because of an incorrect understanding of what was being asked in the first question, it may also come down to differing expectations on the part of assessors (the LLB and JD exams were set and assessed by different instructors).

Intellectual Property Law, Comparative Law, Constitutional Law, Criminal Law, and Administrative Law. In International Business Transactions, ChatGPT did not need to consider any local Hong Kong laws or regulations. However, the assessment was a very practical one, requiring students to analyse a contract and provide specific legal advice. ChatGPT failed to catch many of the issues raised by the contract, and also failed to follow a particular format that students had been taught to use for memo-drafting in class. In Commercial Practice, ChatGPT also had to deal with a very practical exam. Though this question focused on Hong Kong law in particular, no reference was made by ChatGPT to local rules in its answer. However, because the overall topic (a shareholder's agreement) is one that is well-understood across jurisdictions, ChatGPT was able to in the end generate a passable answer.

Intellectual Property Law also featured exclusively problem-based questions that focused on the application of specific Hong Kong laws. And yet, the system was able to refer to general ideas about patent law and list off potential claims raised by the scenario to extent that it received a passing grade. ChatGPT also lost marks as it was unable to rank or judge the merits of any of the claims it was able to spot, despite being asked to do so. The assessment for Comparative Law (2022) consisted of an essay question and hybrid problem-essay questions. The topics were non-local in nature, indicating that ChatGPT likely had more information to draw upon in its answers (though, it was notable that the system was able to write much more about Canada than Thailand). While the essay questions were generally well-written – if rather vague – hybrid questions that asked the system to make particular recommendations about constitutional reform when given a scenario. The system struggled with offering anything beyond general talking points about constitutional reform and received only a C grade.

For the Constitutional Law exam, though the questions were based exclusively on local law, ChatGPT was able to write a reasonable essay on a well-known topic. However, it lacked sufficient detail or reference to the appropriate legal instruments. This was likely because as a 'constitutional' question, it dealt with political developments as much as it did the law and so the system had more material to draw upon. For the problem questions, ChatGPT appeared to benefit from being provided in the exam with the actual text of the legal provisions – this perhaps served to narrow down the range of appropriate answers. The system also appeared to benefit from the hypothetical question being centred around one core legal analysis rather than requiring multiple issues to be spotted. Yet, it still failed to refer to other local legislation or jurisprudence to support its answers, and could apply its understanding of key constitutional provisions in only the most basic way.

The portion of the Criminal Law exam submitted for consideration was an essay question only, for which ChatGPT received the barest of passing grades. The question was a short prompt, but required students to demonstrate understanding of an approach taken by the Hong Kong courts to the question of dishonesty. Because the system failed to refer

to local jurisprudence with any accuracy or detail, the grade was poor. Administrative Law was a similar outcome – a low (but passing) grade on the essay portion of a past exam. The topic was on judicial review of executive power in the context of Hong Kong and COVID19 specifically. While the system could talk in a general sense about limits the courts can place on the executive, it said was unaware of any specific measures the Hong Kong government had taken regarding COVID19 and unaware of any legal challenges to them, and therefore could not properly answer the question. It was interesting to contrast this with the answer to Constitutional Law – because the question itself in that exam explicitly described the measures and the legal instruments that implemented them, ChatGPT was better able to construct a passable answer than it was for Administrative Law.

The rest of the assessments put forward received a failing grade overall, even if certain components were passable: Comparative Law (2021), Company Law, Employment Law, Land Law, Public International Law, Commercial Law, Contract Law, Equity & Trusts, Civil Procedure, and Tort Law. These exams generally failed because they asked very narrow questions about the application of specific Hong Kong legal instruments and required students to apply that knowledge to relatively complicated but entirely fictional (and thus not within the system's training data) scenarios. ChatGPT was also unable to make any kind of value judgment in terms of recommending a particular course of action over another, or being able to gauge the likelihood of success of particular legal arguments before a court, a common element to law exams. Exceptions to this pattern were Comparative Law (2021) and Public International Law. While not having anything to do with local laws, the former failed because it was dependent entirely on access to a text (a specific book assigned as required reading for the course) which was not in the training data. For the latter, even though ChatGPT may have had commentaries and knowledge about international law in its training data, this was not enough to overcome the challenge of accurately spotting and analyzing narrow legal issues in a long, artificial scenario with no textual reference to real-world legal instruments.

Painting with a broad brush, ChatGPT achieved its strongest results on assessments that had the following characteristics: relatively short questions encouraging students to write an essay (reflect, consider, discuss, analyse); a topic that was predominantly non-local in nature; did not require proof of use of specific methodologies taught in the classroom; did not require clear value judgments or predictions of outcomes; and were based on 'real-world' events rather than hypotheticals. Conversely, it achieved its worst results where the assessment required the intake of extremely lengthy background material; where it required synthesis of specific presented facts with legal principles; where the legal knowledge required was hyper-local in nature; where reference to specific classroom activities or materials had to be made; where clear advice or judgment had to be rendered; and where the questions were based on entirely fictional questions. Where those factors point in different directions, ChatGPT can often obtain an 'average' grade.

C *Inadequate Training Data and its Impact*

As noted above, ChatGPT is limited in a temporal sense. It is not constantly updating, and at the time of writing the cut-off date for its training data was (an unknown point in) 2021; it is not be able to answer questions about events that occur after that date.[88] But this did not seem to be an issue for the project – only a single exam (Administrative Law) dealt with real-world events or legal cases from 2022. Instead, the greatest challenge appeared to stem bias in training data towards English language sources drawn from the Anglo-American world as identified in Part 3, above. ChatGPT was far more successful when asked to write about 'general' or 'international' legal concepts. The more local or niche the topic, the more often errors were made it was entirely stumped (or on occasion, the more it flat out made things up like fake court cases). This is a reflection of what it has been trained upon – mostly secondary sources, in English, that exist openly online. It has no access to Hong Kong legal cases or law reports. A similar project run out of a US law school found that ChatGPT could accurately cite and explain a number of US law cases[89] – that is not the case for Hong Kong law.

Instead the system fabricated cases and citations (interestingly, it fabricated mostly believable case names – typically the parties were given Chinese names or those of actual government departments or public bodies in Hong Kong). The answers generated suggest GPT3.5's training data does not include the vast majority of academic books and articles written about Hong Kong legal issues, as they are behind paywalls or other digital limitations. Hong Kong is a tiny jurisdiction in a relative sense, and what open data there is written about its legal issues (such as that included digital media, blogs, message boards, etc) is often in Chinese; there is comparatively little English-language news reports, social media, or expert-written blogs on Hong Kong legal issues as compared to other common law jurisdictions. This therefore impacts the available training data. It was instructive that in the Comparative Law (2022) exam, the system was able to generate a much more detailed response to a prompt about Canadian law than it was Thai law. The conclusion is simple – more information about larger, English-language jurisdictions and the legal issues they face is likely to be included in the underlying training data, and thus ChatGPT is better equipped to answer questions about those jurisdictions.

IV CHATGPT & ACADEMIC INTEGRITY: THE NEW 'CONTRACT CHEATING'?

The misuse of ChatGPT in education was almost immediate. After high school students in one Canadian school were found to be using the tool to complete their work, ChatGPT was banned on school-owned

---

[88] 'Why doesn't ChatGPT know about X?', *OpenAI* (Blogpost) <https://help.openai.com/en/articles/6827058-why-doesn-t-chatgpt-know-about-x>.

[89] Choi (n. 32) 14.

devices and networks.[90] In Australia, Queensland and New South Wales enacted similar bans at all schools in their states.[91] The New York State education department in the United States did the same.[92] Concerns were quickly raised in higher education too: some concluded that ChatGPT would disrupt 'entire tradition [of the undergraduate essay] from the ground up.'[93] Others said they were witnessing the 'death of the college essay in real-time.'[94] One professor declared they would no longer give take-home assignments.[95] Another believed they had detected the use of ChatGPT in 20% of a recent set of submitted assignments.[96]

The fear was, quite simply, that students would use ChatGPT to submit work that was not their own and – crucially – would not be caught. Conventional plagiarism detectors such as 'Turnitin' work by crudely matching the submitted work to text that has already been submitted to it (to catch 'co-operative' cheating) and to text available on the web (to catch 'cut-and-paste' cheating). If portions of text match and there is no appropriate citation, the assignment is flagged for review as containing potentially plagiarising material. This kind of system would not generally catch an AI-generated output because it is unique – even putting the same prompt in twice in a row will not result in an identical output. In response, as this paper was being written OpenAI released a 'detection tool'[97] and Turnitin announced plans to incorporate some kind of AI-detection software into its core product as well.[98]

However, these developments are unlikely to provide a solution to the central dilemma facing educators because they cannot assure assessors with any certainty that simply because a system says a piece

---

[90] Bobby Hristova, 'Some Students Are Using ChatGPT to Cheat – Here's How Schools Are Trying to Stop It', *CBC News* (online, 2 February 2023) <https://www.cbc.ca/news/canada/hamilton/chatgpt-school-cheating-1.6734580>.

[91] Caitlin Casidy, 'Queensland Public Schools to Join NSW in Banning Students from ChatGPT', *The Guardian* (online, 22 January 2023) <https://www.theguardian.com/australia-news/2023/jan/23/queensland-public-schools-to-join-nsw-in-banning-students-from-chatgpt>.

[92] Maya Yang, 'New York City Schools Ban AI Chatbot That Writes Essays and Answers Prompts', *The Guardian* (online, 6 January 2023) <https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt>.

[93] Stephen Marche, 'The College Essay is Dead', *The Atlantic* (online, 7 December 2022) <https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/>.

[94] @corry_wang (Twitter, 1 December 2022 12:44 PM AWST) <https://twitter.com/corry_wang/status/1598176074604507136>.

[95] @Afinetheorem (Twitter, 1 December 2022 6:29 AM AWST) <https://twitter.com/afinetheorem/status/1598081835736891393>.

[96] Caitlin Cassidy, 'Lecturer Detects Bot Use in One-Fifth of Assessments as Concerns Mount Over AI in Exams', *The Guardian* (online, 16 January 2023) <https://www.theguardian.com/australia-news/2023/jan/17/lecturer-detects-bot-use-in-one-fifth-of-assessments-as-concerns-mount-over-ai-in-exams>.

[97] 'New AI Classifer for Indicating AI-Written Text', *OpenAI* (Blogpost, 31 January 2023) <https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/>.

[98] 'Sneak Preview of Turnitin's AI Writing and ChatGPT Detection Capability', *Turnitin* (Blogpost, 13 January 2023) <https://www.turnitin.com/blog/sneak-preview-of-turnitins-ai-writing-and-chatgpt-detection-capability/>.

of work is authentic it in fact is, for two reasons. First, because the 'detectors' are themselves algorithms trying to detect the use of algorithms in outputted text, they too are open to manipulation - immediately after Turnitin's announcement, other services sprang up offering students ways to get around such detection.[99] This is compounded by the second flaw, which also stems from the algorithmic nature of the detectors – they offer only a *probability* that a piece of work contains-generated AI text. OpenAI is not placing a hidden watermark in its ChatGPT outputs that would classify them with 100% certainty as AI-generated. Indeed, it admits that its detector correctly identifies only 26% of AI-outputs as 'likely AI-written', and 9% of the time falsely identifies human-written outputs as AI-generated.[100] Turnitin's AI-detector also generates false-positives,[101] and admits that it cannot 'make a determination of misconduct' and an assessor will need to 'apply [their] professional judgment, knowledge of [their] students, and the specific context surrounding the assignment.'[102]

In high stakes assessments (such as take-home exams worth 100% of a student's final grade – as is often the case in elective courses in law schools), reliance on technology to detect AI-generated outputs may lead to false allegations of plagiarism is a fundamental problem. Unlike more conventional forms of plagiarism – where the allegedly infringing text can simply be directly compared to the original – there can be no 'proof' short of admission by the student. In a high-stakes scenario then, this simply encourages an accused student to deny all allegations. It seems unlikely (and perhaps tortious) that a law school would fail a student and thereby seriously damage their career prospects based only on a 'probability suggestion' from a piece of software that is known to generate false positives. Consequently, at best the existence of new 'AI detectors' might scare off some students from considering submitting an AI-generated work in the first place. They will not solve the fundamental conundrum facing higher education in the age of generative AI.

It is useful to consider the threat generative AI may pose to academic integrity as a 'new and improved' version of 'contract cheating', which Awdry & Newton define as 'the act of students submitting work for academic credit or formative assessment, which they have purchased from an essay mill or other service selling work to

---

99 See eg Daniel Hojris Baek, 'ChatGPT Detector – 10 Tools and How to Get Around Detection', *SEO.ai* (Blogpost, 24 January 2023) <https://seo.ai/blog/chatgpt-detector-tools>.

100 'New AI Classifer for indicating AI-written text', *OpenAI* (Blogpost, 31 January 2023) <https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/>.

101 Geoffrey A. Fowler, 'We Tested a New ChatGPT-Detector for Teachers. It Flagged an Innocent Student', *The Washington Post* (online, 3 April 2023) <https://www.washingtonpost.com/technology/2023/04/01/chatgpt-cheating-detection-turnitin/>.

102 'Understanding False Positives Within Our AI Writing Detection Capabilities', *Turnitin* (Blogpost, 16 March 2023) <https://www.turnitin.com/blog/understanding-false-positives-within-our-ai-writing-detection-capabilities/>.

students.'[103] So-called 'writer for hire' services have long been part of the higher education landscape,[104] but the internet has made purchasing those services easier than ever.[105] Research has shown that contract cheating is widespread, with one 2018 study of Australian universities suggesting that roughly 6% of students admitting to having engaged in contract cheating.[106] It is often of high quality, often undertaken by former students seeking extra money.[107] But what matters here is not the *purchasing* aspect – it is the fact that the work is being completed by another party and passed off by the student as their own.[108] The use of generative AI comfortably fits within this genre, even if the other 'party' is a large-language model run by a multi-billion dollar corporation.

Australia made it an offence to provide or advertise an academic cheating service on a commercial basis in 2019.[109] The same Bill also granted the higher education regulator (Tertiary Education Quality and Standards Agency, or TEQSA) the authority to apply to a court to block websites offering cheating services, which it has repeatedly done.[110] Ireland adopted similar offences in 2019,[111] as did the UK in 2022.[112]

---

[103] Rebecca Awdry and Philip M Newton, 'Staff Views on Commercial Contract Cheating in Higher Education: A Survey Study in Australia and the UK' (2009) 78 *Higher Education* 593, 594.

[104] See eg Sarah Elaine Eaton, 'Contract Cheating in Canada: A Comprehensive Overview' in Sarah Elaine Eaton and Julia Christensen Hughes (eds), *Academic Integrity in Canada* (Springer, 2022); Shawren Singh and Dan Remenyi, 'Plagiarism and Ghostwriting: The Rise in Academic Misconduct' (2016) 112(5-6) *South African Journal of Science*.

[105] In 2015 Australian media revealed the widespread use of cheating services advertised at some of the country's top universities, kick-starting public consternation with the practice: Amy McNeilage & Lisa Visentin, 'Students Enlist MyMaster Website to Write Essays, Assignments', *Sydney Morning Herald* (online, 12 November 2014) <https://www.smh.com.au/education/students-enlist-mymaster-website-to-write-essays-assignments-20141110-11k0xg.html>.

[106] Tracey Bretag et al., 'Contract Cheating: A Survey of Australian University Students' (2018) 44(11) *Studies in Higher Education* 1837, 1840.

[107] Shiva Sivasubramaniam, Kalliopi Kostelidou, and Sharavan Ramachandran, 'A Close Encounter With Ghost-Writers: An Initial Exploration Study on Background, Strategies and Attitudes of Independent Essay Providers' (2016) 12(1) *International Journal for Educational Integrity* 1.

[108] See eg Rowena Harper et al., 'Contract Cheating: A Survey of Australian University Teaching Staff' (2018) 1 *Legal Education Review* 1; Tracey Bretag et al., 'Contract Cheating and Assessment Design: Exploring the Relationship' (2019) 44(5) *Assessment & Evaluation in Higher Education* 676.

[109] Tertiary Education Quality and Standards Agency Amendment (Prohibiting Academic Cheating Services) Bill 2019 (Cth) <https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r6483_aspassed/toc_pdf/19257b01.pdf>.

[110] As of October 2022, TEQSA had successfully applied to block 150 websites targeting Australian students and it said it was working through a 'priority list' of 580. See 'TEQSA Disrupts Access to Another 110 Illegal Academic Cheating Websites', *TEQSA* (Blogpost, 13 October 2022) <https://www.teqsa.gov.au/about-us/news-and-events/latest-news/teqsa-disrupts-access-another-110-illegal-academic-cheating-websites>.

[111] Qualifications and Quality Assurance (Education and Training) (Amendment) Act 2019 <https://www.irishstatutebook.ie/eli/2019/act/32/enacted/en/html>.

[112] Skills and Post-16 Education Act 2022 <https://www.legislation.gov.uk/ukpga/2022/21/enacted>.

But while some regulators have begun to grapple with the problem, they have had limited success due the digital aspect. While TEQSA may have blocked access to a number of websites within Australia, they are all still accessible to any student who knows how to use a VPN. The criminal penalties are severe, but authorities may struggle to apply them to individuals physically located outside their borders and who are running online services hosted overseas, even if they are aimed at domestic students. An international network of higher education regulators has been formed to try and address these issues, but as yet there have been no concrete solutions offered.[113]

Generative-AI assessments are even less likely to be susceptible to criminal prohibition, for several reasons. First, governments are unlikely to want to imprison or fine a student for using an AI tool to help them with coursework (indeed, in their laws aimed at contract cheating all three of Australia, Ireland, and the UK made sure the *user* of the service was not subject to any penalties). Second, the nature of generative AI means that the only alternative is to criminalize the very use of it and while authoritarian states may indeed do so to ensure the 'proper' flow of information is maintained,[114] this is likely a non-starter elsewhere, given the tremendous economic and social benefits generative AI promises.

But while this means approaches to control over contract cheating and generative AI assessments do not completely overlap, the literature about the former is useful in helping us ground our responses to the latter.

V POOLS OF SORROW OR WAVES OF JOY? GENERATIVE AI & THE FUTURE OF ASSESSMENT IN LAW SCHOOLS

The bigger issue is perhaps not what ChatGPT can do now, but what it (and its competitors) will be able to do in a few years. It is reasonable to assume that these kinds of systems will soon have access to real-time data, that the size of their training data will continue expand, that gaps in localized or non-English language knowledge will continue to be filled, and so on.[115] Today's F on a problem-based commercial law

---

[113] John Walshe, 'Global Network Set Up to Stamp Out Contract Cheating in HE', *University World News* (online, 21 October 2022) <https://www.universityworldnews.com/post.php?story=20221021132900222>.

[114] Helen Davidson, ''Political Propaganda': China Clamps Down on Access to ChatGPT', *The Guardian* (online, 23 February 2023) <https://www.theguardian.com/technology/2023/feb/23/china-chatgpt-clamp-down-propaganda>.

[115] Indeed, after GPT4 was released (and after the first draft of this paper had been completed) I re-ran three exams through the system – one of my own and two from colleagues. I had awarded the first of those (Constitutional Law) a C/D grade when answered by GPT3.5 (in other words, the essay portion of the exam was a 'C' level answer while the issue-spotting portion received a 'D' grade). This rose to a B/C under GPT4 – the reasoning was much improved, though there was still inadequate reference to local jurisprudence. GPT3.5 had been completely unable to answer one part of the Comparative Law (2021) exam, but my colleague judged that GPT4 was able to now produce a B-level answer. While GPT3.5 had also failed the Equity &

exam may be a D next year, a C the year after, and A shortly thereafter. In a world where ChatGPT *can* write a top-level law exam answer with regularity, what are we to do?

At the moment, the 'three-hour, open-book' exam is the most common format used at CUHK Law – students sit an in-class exam for three hours, and can bring in any materials they want other than library books and electronic devices (unless they are using a locked-down laptop for an e-exam). It is a near certainty that this format will remain (and in any event, for the time being in Hong Kong in-class exams are mandated by the regulatory body for core courses that lead to legal qualification). An obvious, blanket solution would be to make all assessments of the 'in-class' variety. This has the advantage of being easy to implement, and so long as students are also forbidden from using electronic devices during the exam (or are using 'locked down' systems[116]), then it will be essentially impossible for ChatGPT or other generative AI tools to be leveraged in an unscrupulous fashion.

But take-home exams have notable pedagogical benefits including reduced anxiety for students, allowing for greater synthesis and detail, and allowing for greater application of knowledge to new situations.[117] For the time being, longer research papers seem to be a relatively low-risk form of assessment in the context of ChatGPT. It struggles to produce answers beyond a certain length and above a certain detail. But in any event, catching out a small number of potential 'cheaters' should not be the primary aim of assessment pedagogy. Even in the pre-ChatGPT era, the use of take-home exams persisted despite the potential for increased plagiarism.[118] Likewise, longer research papers or dissertations have increasingly been 'at threat' of plagiarism – the 'writer for hire' services described in Part 4 testify to that.[119] But the intensive dissertation remains an important part of an undergraduate or postgraduate education, at least in law. It would be a disservice to students who are genuinely intellectual curious about a topic to forbid them from writing research papers simply because there is a risk one of their classmates might cheat.

Put simply, universities should not respond to the rise of generative AI by simply trying to ban it. Just as the computer redefined legal practice (and then education) in the 1990s, generative AI looks likely to do the same in the 2020s. Law schools must adapt to this reality not by

---

Trusts exam the assessor judged that the GPT4 answer was much improved, though it relied too heavily in English cases rather than local ones.

116 Though, admittedly, some have argued that locked-down exams are not 100% secure. See Phillip Dawson, 'Five Ways to Hack and Cheat With Bring-Your-Own-Device Electronic Examinations' (2016) 47(4) *British Journal of Education Technology* 592.

117 Corey Johnson et al., 'Assessing and Refining Group Take-Home Exams as Authentic, Effective Learning Experiences' (2015) 44(5) *Journal of College Science Teaching* 61.

118 Arto Hellas, Juho Leinonen, and Petri Ihantola, 'Plagiarism in Take-home Exams: Help-Seeking, Collaboration, and Systematic Cheating' in Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '17) (Association for Computing Machinery, 2017) 238–243 <https://doi.org/10.1145/3059009.3059065>.

119 See above nn 103-108.

attempting to keep it at arm's length, but by embracing it in a manner that is beneficial for our students. The question must quickly shift from 'is AI a threat to traditional legal assessment?' to 'how can legal educators leverage AI for the benefit of students when designing courses?' Answering this properly will take time, and some trial and error. Ultimately, the key to effective AI use as part of legal pedagogy will be instil in students the idea that AI might be wrong, it might be biased, and it almost certainly can be improved upon – and that this will also be the reality of their professional practice. If they cannot improve upon an AI answer, then what do they hope to be paid for?

I briefly sketch out four possible ideas below that might provide a basis for written assessments in the age of AI. The central idea of the first three is to let students learn how to appropriately use (but not blindly trust) answers generated by AI: using AI as a drafting tool, an outlining tool, and an analytical tool. The fourth envisions a world where 'anything goes' – students are directed to answer a question using any technology they wish, but there is a strictly applied grade curve. Following those four ideas, I note that law schools already have an option for assessment that seems largely AI-resistant: clinical legal education.

First, students could be encouraged to use AI to give a 'first draft' answer, and then asked to improve upon that answer manually – in other words, to spot what the AI missed, correct what is misrepresented, or to strengthen its arguments with reference to alternative jurisprudence. Students could ask multiple AI's to perform the same task and be required to explain any divergence and how they intend to reconcile it, or consider whether it reflects an inadequacy in the training data or bias in the code. This could be used in various kinds of assessments, not just the obvious options like a legal writing course. For instance, a common form of assessment question is the provision of a hypothetical scenario for which students must then provide a legal analysis. Students could be asked to have the AI generate one side of the argument, and then they generate the other side. Alternatively, students could be *provided* with a default AI-generated answer as part of the exam, and then asked to improve upon it. In either case, allowing the use of AI in the context of assessable works will probably need to be accompanied by a strict grading curve: how much did Student A improve the AI answer compared to Student B?

Second, AI as an outlining tool. This might be of use in essay-style questions or research papers. Student might be taught how to use appropriate prompts to generate rough plans or outlines of larger papers, and asked to share those outlines with their professor or the rest of the class. As part of that sharing they might be asked to identify possible weaknesses, or make revisions. This might work alongside the first idea – students could use an AI to draft an introduction for them, and share that version along with the final version that they wrote to show the professor how they improved upon it. Leveraging 'AI as an editor' is a related concept – in the context of a longer piece of writing, students could put it through the AI system and ask for improvements, and then

explain why they chose to incorporate or reject those machine-suggested changes.

Third, as an analytical tool. At the risk of generalization, in my experience students rarely read as many complete cases as they should – *if any*. The vast majority are already learning case law primarily by reading a summary they find on Google or from 'handed down' notes rather than immersing themselves in the whole text. Given this (and assuming training data eventually includes complete jurisprudence from multiple jurisdictions) AI could be put to use as an analytical tool. Professors may not like this, but it is better to cede to the reality rather than pretend students are pouring over law reports late into the night. Students could be asked to use AI to draw out key ideas from long cases – particularly new ones. They could use an AI to compare similar cases across jurisdictions, or to track whether minority judgments in one place track majority ones elsewhere. As with the first idea, they could consult multiple AI services and try and reconcile or justify differences in their respective outputs. The short of it is simply to get students to use AI as a tool that aids their learning, rather than being a substitute for learning itself.

Fourth, the 'anything goes' approach. Under this model, generative AI would be completely embraced and students would be encouraged to use it: 'here's an exam, anything goes – but there's strict grade curve.' In other words, equality of arms: the assumption is that (just like in the 'real world') everyone has access to search engines, to a mass of information, and to AI. However, so long as generative AI tools are 'paid-for' commercial enterprises, then issues of equity will likely arise unless universities are able to negotiate open access to the tools for the entire student body. But even if equal access to the tools could be guaranteed, the 'anything goes' approach would still require some systemic changes in pedagogy. Ensuring equal arms is one thing, but ensuring everyone is a good shot is another. Without some experience with the first three approaches, students would likely struggle if thrown into the fourth.

But while written assessments have long been the dominant in law schools, they are not the exclusive option. 'Clinical' legal education seeks to allow students to put into practice their knowledge and skills in real-world settings.[120] A subset of this approach uses 'simulated clients' to help students build practical skills at all stages of the learning process (this latter concept is borrowed from medical schools).[121] How

[120] See eg Elliott S. Milstein, 'Clinical Legal Education in the United States: In-House Clinics, Externships, and Simulations' (2001) 51(3) *Journal of Legal Education* 375; Neil J. Dilloff, 'Law School Training: Bridging the Gap between Legal Education and the Practice of Law' (2013) 24(2) *Stanford Law & Policy Review* 425; Mark Spiegel, 'Theory and Practice in Legal Education: An Essay on Clinical Education' (1987) 34(3) *UCLA Law Review* 577; Margaret Martin Barry, Jon C. Dubin, and Peter A. Joy, 'Clinical Education for This Millennium: The Third Wave' (2000) 7(1) *Clinical Law Review* 1.

[121] See eg Ian Weinstein, 'Testing Multiple Intelligences: Comparing Evaluation by Simulation and Written Exam' (2001) 8(1) *Clinical Law Review* 247; Fran Wasoff & R. Emerson Dobash, 'Simulated Clients in Natural Settings: Constructing a Client to Study Professional Practice' (1992) 26(2) *Sociology* 333; Deborah Maranville,

precisely students would use generative AI in clinical education is a separate question that is more closely tied to how lawyers will use it themselves. The point here is that assessment in law school need not only be defined as written outputs – there is already a vast body of literature that shows how clinical education with either real or simulated clients benefits the learning process.

But all these options require intensive resource commitments. Having academic staff using generative AI as part of the learning process in the three ways I described above is not plausible in a lecture of 80 or more students and zero teaching assistants – the model of a first-year course not only at CUHK Law but at many law schools. Expanded clinical education programmes require a deep (and expensive) commitment on the part of both a law school and the relevant professional regulators to seriously re-think how legal education is delivered. Because of the scope of the challenge generative AI poses to traditional teaching practices, it is unfortunately foreseeable that less well-resourced or less well-led law schools will instead simply mandate 'in-class only' exams. Without appropriate funding, effective leadership, and regulatory support, they will attempt to continue delivering the material in otherwise the same fashion as has traditionally been done.

This would be a mistake. Generative AI tools like ChatGPT are poised to revamp all kinds of knowledge-work. Legal practice and legal education (and higher education generally) will not be immune from these changes. While this project has shown that for the moment traditional forms of legal assessment tend to not lend themselves well to high quality answers written by ChatGPT, it would be foolish to assume that that will be true in the future. While professors in all fields would do well to consider the form of their current assessments in light of the emergence of these new tools, the nature of the beast means the range of questions that can stump generative AI will constantly shrink over time.

But institutions of higher education around the world need to ask themselves questions not just about the viability of their assessment schemes in the short term, but the viability of the pedagogical models more broadly in the long term. Just as law students will have to accept that they will only be hired into legal practice if they can demonstrate some kind of skill that goes *beyond* what an AI can do, professors will have to demonstrate that they can offer students something beyond the same old material delivered and examined in the same old way. To do

---

'Passion, Context, and Lawyering Skills: Choosing Among Simulated and Real Clinical Experiences' (2000) 7(1) *Clinical Law Review* 123; Patricia A. Hollander, 'The Simulated Law Firm and Other Contemporary Law Simulations' (1977) 29(3) *Journal of Legal Education* 311; Lawrence M. Grosberg, 'Medical Education Again Provides a Model for Law Schools: The Standardized Patient Becomes the Standard Client' (2001) 51(2) *Journal of Legal Education* 212; J.M. Feinman, 'Simulations: An Introduction' (1995) 45(4) *Journal of Legal Education* 469; Karl S. Okamoto, 'Learning and Learning-to-Learn by Doing: Simulating Corporate Practice in Law School' (1995) 45(4) *Journal of Legal Education* 498; Carol R. Goforth, 'Use of Simulations and Client-Based Exercises in the Basic Course' (2000) 34(2) *Georgia Law Review* 851.

anything else would be a disservice to students; the AI tide will not go back out. By accepting and incorporating AI into both teaching and assessment, universities are more likely to better prepare students for a world of work that will increasingly rely on such tools as part of knowledge generation.

VI APPENDIX

Below is the list of exams from CUHK Law to which ChatGPT was subjected, an estimated grade provided by an expert assessor, along with a brief explanation of the requirements of the exam and commentary by that assessor. They are organized from highest grade achieved to lowest.

Course: Jurisprudence
Programme: JD
Exam Style: Essay questions
Grade: B+/A-

This exam was composed of two questions. The first was a short prompt asking students to provide a defence of utilitarianism. The second asked students to reflect on a quote from Nigel Simmonds on the applicability of utilitarianism as both a theory of personal morality and as a guide for government policy.

The assessor judged that ChatGPT's answer to the first question was a mid-range B, while the answer to the second would have received a A-, possibly even an A. They were of the view that the answer to the second question was mirroring the sophistication of the question. They suggested that had the first answer had been more creative it likely would have received in the A range as well. ChatGPT marked a turning point in assessment for this assessor: after seeing these results they were of the view that all future assessments in law school would have to be 'in-class', and that courses based on dissertations would have to be scrapped. This was perhaps the most pessimistic view the project received as feedback: 'I fear this is the new reality.'

Course: International Environmental Law
Programme: JD/LLM
Exam Style: Essay questions
Grade: B+/A-

The exam was composed of 3 essay style questions, each of which was a short prompt asking to students to critically analyse or discuss some aspect of the course. The first was about whether sovereign claims to natural resources gives states unconstrained rights to exploit fossil fuel reserves; the second required an analysis of how the Paris Agreement aims to limit global warming; the third asked whether world trade law is a force for good or an obstacle in implementing climate change mitigation policies.

The assessor deemed the answers all 'impressive… generally solid and well argued.' The essays considered appropriate pros and cons, and correctly linked various legal concepts. While there were some inaccuracies, none were fatal. The most obvious weakness was a lack of references to cases, literature, or any case material.

Course: International Tax Law
Programme: JD/LLM
Exam style: Essay questions

Grade: B+/A-

This exam consisted of three essay questions. The first required students discuss a statement from the OECD's 2013 Action Plan on Base Erosion and Profit Shifting (BEPS). The second required students to analyse an academic quote on the cause 'base cyberisation', and do so by highlighting legal and regulatory developments on both the national and international levels. The third required students to discuss the importance of offshore indirect transfers to developing countries.

The assessor described the answers as well structured, correctly stating the key issues, and using appropriate language. Though the answers lacked some of the necessary details and included some irrelevant examples, on balance the answers were 'generally impressive'. Overall, the assessor found the ability of ChatGPT to be 'worth great attention and concern.'

Course: Hong Kong Legal System
Programme: JD
Exam style: Essay questions
Grade: B/B+

This exam was composed of two questions. The first provided students with the text of Art. 17 of Hong Kong's Basic Law which vests it with legislative power, and asked for a consideration of the provision with respect to Hong Kong's law-making process and its exercise of a high degree of autonomy. The second asked students to compare and contrast the interpretive powers of the courts of Hong Kong and those of the Standing Committee of the National People's Congress (NPCSC).

The assessor deemed the answer to the first question to be substantially better than that to the second. They found that ChatGPT appeared to rely on the same information to answer both, and this would be judged poorly in a real-world context. The second answer failed to catch the focus on Interpretations of the Basic Law by the NPCSC (a critical issue in Hong Kong law), and also failed to explain the interpretive tools used by the local courts at a statutory level. The system also made up non-existent cases (eg *Tam Yiu-chung v. HKSAR*) and a non-existent Interpretation by the NPCSC ('the 1999 Interpretation of Art. 23'). Despite this, the assessor remarked that 'you could've fooled me and said that one of our students wrote this… the answers are good enough but not great, [and] that makes it believable.'

Course: Legal System of the PRC
Programme: LLB
Exam style: Essay questions
Grade: B

Two questions were submitted for consideration. The first asked students to consider a comment on traditional Chinese social order and provide a rationale for agreement or disagreement by making reference to two films: the Story of Qiuju and The Accused Uncle Shangang. The second question asked students to consider the feasibility of the

introduction of legislative review with reference to the Seed Case and the Case of Sun Zhigang.

The assessor concluded both were deserving of a B grade. While they found there to be some errors in the systems description of the relevant films and cases, the answer reflected a decent general understanding, if not being in depth or critical enough to stand out. The assessor found this on the one hand to be 'quite remarkable' given the subject matter, but also not too concerning from a pedagogical perspective if the system could only meet a barely average level of quality.

Course: Competition Law
Programme: JD/LLM
Exam Style: Mixed problem and essay questions
Grade: B/B-

The exam was composed of three questions. The first was a problem style question in which students were presented with a hypothetical scenario involving the Competition Commission commencing proceedings against a distributor of imported foodstuffs for alleged resale price management. Students were assigned the role of defence counsel and asked to identify all possible non-procedural defence strategies. The second question was a short essay question asking students to consider and critique a quote from the chair of the Hong Kong Competition Commission on holding parent companies liable for contraventions committed by subsidiaries. The third question asked students to consider the implications of the different wording as between Hong Kong and European Union law when defining companies holding a certain level of market power that might trigger scrutiny.

The assessor found the answer to the problem question the weakest, though 'not terrible', giving it a B- grade. It was written smoothly, but featured an incorrect description of a portion of the relevant law and cited a non-existent case ('*Hong Kong Competition Commission v Cathay Pacific Airways Limited*'). The essay questions were both judged to result in B level answers. The assessor's conclusion was that the results were 'scary', given that the system would inevitably improve over time.

Course: International Business Transactions
Programme: LLM
Exam style: Problem question
Grade: B-/C+

Students were provided with a hypothetical scenario involving a joint venture between companies located in two jurisdictions planning to open laboratory in a third. They were also provided with excerpts of a draft contract related to this venture, and asked to produce a memorandum in response to a series of requests from the Chief

Financial Officer of one of the companies, including one for general comments on the draft.

The assessor judged that the general parts of the memo were acceptable, but a number of specific issues and problems contained within the draft contract were missed. ChatGPT also failed to follow a particular form of memorandum that students had been told to use in class.[122] While concluding that the answer was likely of a high C or possibly low B grade, the assessor noted that it 'was much better than [they] expected.'

Course: Commercial Practice
Programme: PCLL
Exam style: Problem questions
Grade: B-/C+

The portion of the exam submitted for consideration consisted of a hypothetical scenario involving planned restructuring of a company and the associated creation of shareholders' agreement between three partners. Students had to consider the scenario and answer two questions related to how best to protect the position of one of the partners through the content of a shareholders' agreement.

The assessor judged that the answers were written in an overly general fashion, but nonetheless were able to identify some of the relevant issues. No reference to any Hong Kong-specific law was made, but the answers certainly reached a passing grade – probably a low B or a high C. In their view, the system was strong enough that changes in teaching style were probably necessary: 'less diligent students [might] simply rely on [ChatGPT] without the need to prepare for tutorials.'

Course: Jurisprudence
Programme: LLB
Exam style: Essay questions
Grade: B/D

This exam was composed of three questions. The first asked students to whether virtue ethics theory had any method to justify what is morally right or wrong, and to compare virtue ethics theory with utilitarianism and deontology when answering. The second to compare Finnis' and Hart's notions of natural law. The third to consider the extent to which a Dworkinian judge would take into popular morality when deciding cases.

The assessor judged the first answer to receive a D grade, on the ground that ChatGPT answered a different question than it was asked. The system appeared to only consider the second part of the question, and instead spent too long summarizing each of their theories and highlighting their similarities and differences. The second and third

[122] Presumably a student using ChatGPT would know this and be able to redraft the output as appropriate.

questions received a low B and a high B, respectively. The second question also spent too long summarizing theories rather than directly answering and was too superficial overall. The third question was better; the assessor judged that thought content was not particularly deep, ChatGPT answered the question 'dead on' and accurately.

Course: Intellectual Property Law
Programme: LLB
Exam style: Problem questions
Grade: C

The portion of the exam that was submitted for consideration was a hypothetical situation involving patent infringement claims regarding two drugs made available for sufferers of acid indigestion. Students had to identify the most likely grounds for a successful claim of patent infringement and then the best way to defend against that claim, as well as judge its likelihood of success.

The assessor concluded that ChatGPT had 'skirted the main issues' in the scenario, though what was presented was well organized and touched on enough relevant points that it deserved a C grade. The system was able to discuss general principles of patent law and even connect some of them to the facts presented, but was unable to 'drive home' an argument one way or another. The assessor found that while the system was able to list off potential claims or arguments, it was unable to rank or judge the merits of any of them as more or less plausible. In their view, 'it generally knows what to say, but does not know the value of what it says.' Despite these obvious weaknesses, they were impressed overall: 'I had expected ChatGPT to flounder on the answer. It did not.'

Course: Comparative Law (2022)
Programme: JD/LLM
Exam Style: Mixed problem and essay questions
Grade: B-/C

This exam contained three parts. The first was an essay question that required students to reflect on Ugo Mattei's claims about taxonomies of law by using Hong Kong as a case study. The second was a problem question that asked students to recommend amendments to a particular constitution that they had been assigned in the course,[123] with detailed explanations as to why they had been chosen. The third was an hybrid hypothetical-essay question that asked students to give advice to the drafters of a constitution about whether they should allow a constitutional court to adjudicate fundamental rights, and if so, whether they should adopt a strong or weak form of judicial review.

For the first question, the assessor found it surprising that the failed to correctly describe the taxonomy (which they considered the easy

[123] ChatGPT was asked to answer the question twice – once as though they had been assigned to study the Canadian Charter of Rights & Freedoms, and once as though they had been assigned the Constitution of Thailand.

aspect) but was able to better use Hong Kong as a case study to critique it (the harder aspect). Hong Kong's legal system was described accurately, and the structure of the answer was clear. While the assessor said they would have expected more detail and nuance, it was at the same level of many actual student answers to the question. It would have received a B- grade.

For question two, the assessor found that the answer to be of a B-level, because it was able identify some potential amendments that made general sense, and cite case studies and some literature. However, closer scrutiny revealed that two of the articles referred to was made up and another while genuine did not accurately summarize the author's position. Moreover, the answer provided when assigned to talk about Canada was significantly longer than when talking about Thailand, indicating that the amount and quality of the training data varies across jurisdictions. No literature or case studies (even invented ones) were provided in the answer for Thailand. Question three was deemed to be of a C level, as the arguments were underdeveloped and there was little analysis 'beyond truisms about the role of constitutions.'

Course: Administrative Law
Programme: LLB
Exam style: Essay question
Grade: D

The portion of the exam submitted was an essay question. It presented students with a quote from two professors (neither from Hong Kong) regarding global constraints on executive power during the COVID19 pandemic, suggesting courts have largely prioritized public health over other considerations. The question then asked students to consider to what extent the approach of the courts in Hong Kong reflected this view.

The system struggled with the initial prompts for this question, saying it had no knowledge of the responses of how the courts dealt with COVID19. After rephrasing the prompt to be more direct ('comment on the approach of the courts to challenges brought regarding COVID19 policies'), ChatGPT provided an answer. The assessor deemed it to be of a D quality, noting that while it was clearly written it was too brief and superficial, without any consideration of the relevant jurisprudence and different judicial approaches reflected therein. As with some other exams, it invented non-existent cases and citations in support of its answers.

Course: Constitutional Law
Programme: JD
Exam style: Mixed problem & essay questions
Grade: C/D

The exam consisted of two parts. The first part was an essay question asking students to critically consider the composition and selection process of the Court of Final Appeal of Hong Kong. The

second part was composed of two hypothetical questions related to real-world laws that Hong Kong had introduced in early 2020 in its effort to combat COVID19 – one related to mandatory quarantines for most arriving passengers, and one related to a limit in public group gatherings to no more than four people. Students (and ChatGPT) were provided with the text of these laws and other relevant regulations in the exam, along with stories about two fictional clients to whom they then had to provide legal advice about a constitutional challenge to the laws. A follow-up question asked students to also consider what remedies they would approach the court for should the challenge be successful.

The essay question was well-structured, written clearly, and did cover several salient points about the composition and selection methodology of the Court. However, the answer did not go into sufficient detail nor did it make any reference whatsoever to the relevant laws that govern the way the Court operates. This question received a 'C' grade, but with the insertion of just a few citations to materials that were studied in the course might well have received a 'B'. The problem questions regarding COVID regulations were answered less successfully. Though the conclusions reached in the answers to the problem questions may well have been correct at a basic level, they were not supported with an appropriate legal analysis. The answers did highlight the importance of the question of proportionality to any constitutional challenge, but did not explain the necessary four-step test nor apply the facts provided to each step before advancing a conclusion. There was no reference to any jurisprudence covered in the course, which was necessary to achieve a reasonable grade. The short answers about possible remedies were well-written and accurate, but only accounted for a few marks and so were not enough to save the problem questions from receiving a D or possibly C- grade, if the assessor were feeling generous.

Course: Criminal Law
Programme: LLB
Exam style: essay question
Grade: D

The essay portion of an undergraduate-level criminal law exam was submitted for consideration. It asked students whether the approach taken by the Hong Kong courts to determine whether a person had acted dishonestly remained 'fit for purpose'. Interestingly, the first time this question was put though ChatGPT the system generated an answer that explained (the entirely fictional) 'fit for purpose' test as applied in the courts of Hong Kong. It referred to made-up cases describing the application of this made-up test. The answer was so obviously wrong that I assumed even the most disinterested student would notice, and altered the prompt slightly to try and generate a more accurate response.

The assessor judged the essay of relatively poor quality, reminding them of an encyclopaedia summary. It was often repetitive and made insufficient reference to both local jurisprudence as well as relevant common law elsewhere as a comparator. The lack of citations were

problematic and there was no evidence of having completed the readings assigned in the course, which would have helped the student answer the question. The assessor concluded by noting that while the answer was poor, 'it wouldn't fail, and in that sense should be a concern.'

Course: Civil Procedure
Programme: JD
Exam style: Problem questions
Grade: D/F

This exam consisted of a long hypothetical scenario regarding litigation over a defective consumer product allegedly sold by a manufacturer aware of the product's flaws, and asked students a series of procedural questions related to it.

The assessor judged that answers to the hypothetical had a wide range of scores. Two failed, one had a bare pass, and one might even have been an A. They reflected that sometimes the system seemed to have a good understanding of the issues and could even appear to apply facts properly, however at other times it missed very basic components of the law and occasionally made up entirely 'rubbish' answers. While the answers were a mixed bag, the assessor considered that the system had done better than they anticipated and were pessimistic: 'This exercise did surprise, and more so, troubled me. Skynet has just arrived and it will learn quickly.'

Course: Tort Law
Programme: LLB
Exam style: Mixed problem & essay questions
Grade: D/F

This exam consisted of two parts. The first part was a long hypothetical scenario, followed by a direction to students to 'advise all parties suffering loss, damage, or injuries on their possible tort claims.' The second part was composed of three essay questions, from which students had to choose one to answer. The first essay question asked students to discuss a textbook quote on the concept of the duty of care. The second asked students to discuss a textboook quote on the nature of fault. The third asked students to discuss how the decision in *Fairchild v Glenhaven Funeral Services* had changed the law regarding proof of factual causation. In all three questions, students were directed to make reference to decided cases and legal principles.

The assessor judged the first part (consisting of the hypothetical scenario) to be a fail. The answer failed to consider a number of important possible claims that arose from the facts provided, and of the claims that were properly identified they were not discussed in any kind of detail. No jurisprudence was provided to support the claims made, even though direction was specifically given to refer to cases where possible. The answers to the essay questions that made up the second part of the exam were judged to be better, though still of relatively poor

quality – each receiving a passing grade, but only just. Each question accurately set out the basic issues raised, but struggled to properly articulate policy considerations in meaningful detail. In the first essay question an explanation of *Spartan Steel* was incorrect, and it also cited *Heron* which was not relevant. In the second, ChatGPT did not consider any cases at all but the whole question was centred on a discussion of the ways in which courts use fault as a control mechanism. In the third, the answer failed to discuss any problems created by *Fairchild* or any developments in the law subsequent to it.

Course: Public International Law
Programme: JD/LLM
Exam Style: Problem questions
Grade: D/F

This exam involved an extremely long fact pattern involving two hypothetical states involved in various disputes regarding both people and territory, along with threats of escalation. For the first question, students were given a very general prompt to 'analyse the international legal issues resulting from this scenario'. The second question then asked students to predict what the content of a request for an Advisory Opinion by the UN General Assembly to the International Court of Justice would contain, and what the resulting Opinion would likely say.

The length of the scenario (3 pages) was a challenge; ChatGPT struggled with the format and so the entire situation was presented to it in separate chunks of text. It was of little use – the assessor deemed the answer to the first question a failure, calling it 'bizarre' and 'nonsensical'. While ChatGPT was able to identify some general points about treaty violations and legal principles for deciding territorial disputes, it could not accurately apply them and missed various key issues related to state responsibility and the doctrine of sources. The second answer was considered a low D at best, and also described by the assessor as recommending a 'bizarre and unrealistic' approach to international law.

Course: Comparative Law (2021)
Programme: JD
Exam style: Essay question
Grade: F

One portion of a Comparative Law exam from an earlier year was also submitted by a different professor than had submitted the 2022 version. It was an essay question that asked students to argue whether or not the course textbook was Eurocentric, and asked them to refer to specific examples within the first four chapters of that book. ChatGPT had no access to the text of the book in question, and thus could only provide a brief explanation of the meaning of Eurocentrism generally. It was a clear failing grade.

Course: Company Law
Programme: PCLL
Exam Style: Problem questions
Grade: F

This exam was composed of six questions. The first asked students to explain whether an individual in a hypothetical scenario would be considered either a director or shareholder of a subsidiary company due to their role in a parent company. The second question was a long 'fact-pattern' style scenario about the listing of shares and a series of sub-questions about the effects in law and practice on that listing. The third question was a shorter fact pattern style question about a hypothetical joint venture and a question about whether it would fall under the meaning of a 'connected transaction' under the Companies Ordinance. The fourth question provided students with a scenario about a company overstating its profits repeatedly, and students were asked to explain the possible liability for the accountants involved. The final question was a long hypothetical scenario about possible criminal liability under the Securities and Futures Ordinance for insider dealing.

The assessor judged that overall the exam would fail, finding that ChatGPT did poorly when required to cite specific Hong Kong regulations or legislative provisions. It did better however when explaining general advantages and disadvantages associated with a particular fundraising method. It was successful in identifying the relevant offences raised in the scenarios, however it struggled to apply legal rules or precedents to facts provided to students. Naturally, it was also unable to refer to specific points emphasized in lectures as they did not form part of its training data.

Course: Employment Law
Programme: JD
Exam Style: Problem questions
Grade: F

The exam consisted of a two-page hypothetical scenario involving workers at a restaurant, on the job injuries, threats, eventual terminations of employment, and a list of specific claims made by the employees. Students were also provided with three employment contracts related to the parties. Students were asked to 'discuss any employment law issues arising' from the scenario and materials. It was a challenge to provide all the information contained within the exam to ChatGPT. As with Public International Law, the attempted solution was to break it into chunks. The system was first asked to consider the hypothetical scenario, then separate prompts were used to try and teach about the content of the individual contracts.

The assessor judged the resulting answers to be of very low quality, critiquing the answer as overly brief for a three-hour exam, superficial, and lacking in reference to specific Hong Kong jurisprudence and legislation. ChatGPT's understanding of the law, they said, 'read like a quick Google search.' The assessor also remarked that ChatGPT had appeared to confuse some basic elements of the contracts including the

wage rate (this suggests the multiple prompts used to try and 'teach' ChatGPT about the scenario were not entirely successful) and was unable to recognize that they breached the Minimum Wage Ordinance. With that said, the assessor judged that the system was able to spot the 'big issues', and if it had the ability to expand upon its answers by including accurate references to local legislation and jurisprudence, it likely would have been a C- level answer.

Course: Land Law
Programme: LLB
Exam Style: Mixed problem and essay questions
Grade: F

One question from a previous undergraduate Land Law exam was provided for assessment. It involved a short hypothetical, describing a married couple purchasing a flat through a mortgage, with an uneven monetary contribution between the couple and title in the name of only one of them. Problem questions related to the existence of beneficial interests and priority of rights, while a short essay question asked students to evaluate the role Lady Hale's judgment in *Stack v Dowden* has played in the development of the common law constructive trust.

On an initial skim of the answer, the assessor judged it might score a C overall, however following a deeper review, they revised that to a failing grade. The assessor did note that the short essay question was answered substantially better than the problem questions. For the latter, they found that while the answers were plausible, they lacked reference to any appropriate legal authority in Hong Kong and thus did not show the appropriate level of knowledge needed to pass this portion of the exam.

Course: Contract Law
Programme: JD
Exam style: mixed problem and essay questions
Grade: F

Two questions were submitted for consideration. The first was a hypothetical scenario involving reliance by one party on information provided by another leading them to purchase shares. Students had to identify damages that could be recovered as a result of facts contained within that scenario. The second question asked students to reflect on a quote taken from a Hong Kong case about the application of English contract law locally.

The assessor judged that both answers would fail, though only just. The problem question failed to accurately explain fraudulent misrepresentation or identify how the legal principles would apply to different statements provided within the hypothetical scenario. It also failed to consider the application of Hong Kong statutory law directly on point. ChatGPT also missed certain 'important points' and failed to give consideration to the appropriate quantum of damages that might be available. For the essay question, ChatGPT wrote a well-structured

answer but made basic errors in substance including misidentifying legislation and the leading authorities. Much of the detail that was included was deemed 'irrelevant'. Overall, the assessor noted it was 'a close run thing', and that they were concerned that already ChatGPT came up with answers that were better than those of 'a non-trivial minority of students.'

Course: Commercial Law
Programme: LLB
Exam style: Problem questions
Grade: F

The exam consisted of two hypothetical scenarios, each with a series of short questions for students to answer based upon the scenarios. The assessor concluded the answers were 'totally off' and even if they were able to correctly identify the relevant area of law, the explanations were incorrect. In their view, the system still had a 'very, very long way to go to pass the LLB… let alone replacing us!'

As with some other exams, ChatGPT 'supported' its answers for this exam by referring to several local cases. Upon review, none of these cases were real. The system even created fake citations for them, such as *C.K. Cheung v. Tai Hing Cotton Mill Ltd* (HKCFA 16/1993). Not only does the case not exist, the claimed citation is an impossibility – the Hong Kong Court of Final Appeal did not come into existence until 1997.

Course: Equity & Trusts
Programme: JD
Exam style: Mixed essay and problem questions
Grade: F

The portion of the exam that was submitted for this course was an essay question that provided students with a quote from a textbook commenting on *Re Goldcorp Exchange*. Students were then asked to discuss that statement in light of competing theories about *Re Goldcorp Exchange* and *Hunter v Moss*. A voluntary mid-term question was also submitted for consideration – it was a hypothetical situation involving a will with various provisions. Students had to consider their validity.

The assessor found that the essay answer was 'quite impressive in general style', but that ChatGPT appeared confused about the facts of the relevant cases and misidentified the major issue before the court in *Hunter*. With that said, they believed the answer would pass if the system were able to provide a little more detail of the actual cases and expected that with time, 'I am sure the AI can develop to deal with this.' The answers to the problem question were completely wrong, and failed to cite any case law so it was hard for the assessor to find marks that might allow a 'generous pass'.