Bond University

Legal Education Review

Volume 34 Issue 1

2024

Persuasive Legal Writing Using Large Language Models

Damian Curran
University of Melbourne

Inbar Levy
University of Melbourne

Meladel Mistica
University of Melbourne

Eduard Hovy
University of Melbourne

PERSUASIVE LEGAL WRITING USING LARGE LANGUAGE MODELS

DAMIAN CURRAN,[*] INBAR LEVY,[~] MELADEL MISTICA,[^]
EDUARD HOVY[#]

[*] School of Computing and Information Systems, University of Melbourne.

[~] Melbourne Law School, University of Melbourne.

[^] Melbourne Data Analytics Platform, University of Melbourne.

[#] Melbourne Connect, University of Melbourne.

I INTRODUCTION

Lawyers write to persuade. Their writing can take many forms, from a written submission to a judge, correspondence between solicitors, a judgment, or policy statements written by legal advocacy groups. It is designed to target and persuade its audience, be they a judge, an opposing solicitor or the general public. Call it written advocacy, legal argumentation or persuasive legal writing. In any guise, it is 'essential to the practice of law.'[1]

Budding lawyers hone their persuasive writing skills during their studies. Recently, however, Large Language Models ('LLMs') have proved adept at a broad range of language tasks, including many of the elements essential to persuasive legal writing. The potential for law students to use these tools to produce essays and exam answers could undermine the utility of many existing forms of assessment.

Powerful LLMs are already widely available for public, and often free, use. ChatGPT and its plug-ins, along with other similar architectures, were deployed at pace in 2023 and are likely to proliferate. They are already used by many students.[2]

LLMs also present a potential boon for the legal industry.[3] Whilst they may not be ready immediately for unedited use in the courtroom,[4] a robust system capable of producing persuasive legal writing has potentially massive application in the production of draft briefs, draft opinions and judgments, legal judgment prediction, and other day-to-day legal work.

This research first establishes the constituent elements of persuasive legal writing and reviews the available literature on LLM competence in each area. We conducted an experiment comparing the grades received in a law school examination of essays produced by law students, against essays produced by an LLM on the same task. Four essays across two essay topics from the exam of Legal Theory, a graduate law class at the University of Melbourne, were produced by

[1] Michael R Smith, *Advanced Legal Writing - Theories and Strategies in Persuasive Writing* (Aspen Law and Business, 3rd ed, 2013).

[2] A survey by BestColleges in early 2023 of 1,000 university students found that 22% of all respondents had used AI tools, like ChatGPT, on assignments or exams. (See Lyss Welding, 'Half of College Students Say Using AI on Schoolwork Is Cheating or Plagiarism', *Best Colleges* (17 March 2023) <https://www.bestcolleges.com/research/college-students-ai-tools-survey/>.) A study by Valova et al of 102 university and high school students found that over 49% of respondents had successfully used ChatGPT for their academic work, and a further 21% regularly used it, regardless of the results. (See Irena Valova, Tsvetelina Mladenova and Gabriel Kanev, 'Students' Perception of ChatGPT Usage in Education.' (2024) 15(1) *International Journal of Advanced Computer Science & Applications* 466).

[3] McKinsey & Company, a consultancy, estimates that generative-AI (of which LLMs are one example), could generate $180 to $260 billion in annual value in Risk and Legal business functions. See Michael Chui et al, 'The Economic Potential of Generative AI' [2023] *McKinsey & Company*.

[4] For example, Benjamin Weiser and Nate Schweber, 'The ChatGPT Lawyer Explains Himself', *The New York Times* (Online Article, 8 June 2023) <https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html>.

the LLM.[5] All essays were graded by several experienced graders. Both the student and the LLM essays were deidentified and the graders were not told in advance about the provenance of any of the documents. Nor were the graders told that some of the essays were artificially produced. The differences in performance between the artificially generated essays and the honors students were measured.

The LLM output, whilst still on average receiving a passing grade, performed worse than the students. As expected, the output was relatively well structured and argued. Given the well-known tendency of LLMs to hallucinate, it was unsurprising that it did not exhibit accurate knowledge. But despite suggestions in some work that LLMs could exhibit creativity, there was no elevated performance in critical analysis and originality compared to other criteria. This muted performance may be the result of the prompt engineering that was conducted to curtail hallucinations. The grader feedback on the LLM essays also showed greater negative sentiment than the comments for the student essays.

This paper is structured as follows. It first explains how LLMs are trained and why they may be suitable for the production of persuasive legal writing. A survey of the constituent elements of persuasive legal writing and a review of the available literature on an LLM's competence in each area is then conducted. The mixed results of recent experiments in which LLMs have been used to generate similar legal and essay-style writing, and the absence of literature using these tools to generate long-form written content, are noted. The methodology, including the system and prompts used to generate long-form text from the LLM, details of the human essays being used as a benchmark and the grading rubric are discussed, followed by the results. The paper closes with a reflection on the prompt engineering process, and broader observations about LLM bias and the implications of this technology on the legal profession.

II LARGE LANGUAGE MODELS

Language models, at their simplest, predict the likelihood of a sequence of words.[6] This ability enables a diverse range of downstream language tasks. 'Pretrained' language models built upon the transformer architecture with self-attention mechanisms have proliferated since the late-2010's.[7] Their development resulted in notable performance gains across many natural language task benchmarks.[8] Research found that

[5] The University of Melbourne, ‘The University of Melbourne, Legal Theory (LAWS50031) Handbook Entry’ <https://handbook.unimelb.edu.au/subjects/laws50031>.

[6] See Wayne Xin Zhao et al, ‘A Survey of Large Language Models’ [2023] arXiv.

[7] The 'transformer' model based on attention mechanisms was first proposed in Ashish Vaswani et al, ‘Attention Is All You Need’ (Conference Paper, 31st Conference on Neural Infomration Processing Systems, 12 June 2017).

[8] See Jacob Devlin et al, ‘BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding’, ed Jill Burstein, Christy Doran and Thamar Solorio [2019] 1 (Long and Short Papers) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

making these models larger resulted in predictable performance gains, including surprising 'emergent' abilities.[9] These scaled pretrained language models were coined 'large' language models. The most recent LLMs, called generative LLMs or generative chatbots, include the ability to produce language in response to an input query. Since OpenAI's ChatGPT was released in Nov 2022, many similar tools have been developed.[10]

This study, conducted during 2023, used OpenAI's then latest publicly accessible LLM, GPT-4.[11] An OpenAI model was chosen because they underpin ChatGPT, the consumer product that received much public attention in late-2022 and appeared to be most closely associated in the public consciousness with the recent resurgence in interest in AI's language capabilities. The specific model, GPT-4, was used in other recent studies, allowing for comparison of abilities across similar tasks.[12]

Due to 'the competitive landscape and safety considerations,' the architecture and training set of GPT-4 are not publicly available.[13] Nonetheless, some relevant information about its architecture and training data can be gleaned from its technical report and system card and papers on earlier GPT iterations.[14] GPT-4 is trained in two stages. It is first trained to predict blanked-out words from a large dataset of text.[15] It is in this stage that its underlying 'knowledge' is encoded into its parameters. Its training data is likely to include at least the massive corpus used when training GPT-3 and, relevantly to this study, the Wikipedia entries on well-known legal theorists and their ideas. It is also likely to include many more related documents from the web, such as the theorist's original texts, student essays, blogs, texts and academic articles on legal theory.

GPT-4 is then fine-tuned for dialogue in a process known as 'reinforcement learning from human-feedback' ('RLHF').[16] In this

Language Technologies 4171. See also Alec Radford et al, 'Language Models Are Unsupervised Multitask Learners' (2019) 1(8) OpenAI blog 9.

[9] See Jason Wei et al, 'Emergent Abilities of Large Language Models' [2022] Transactions on Machine Learning Research <https://openreview.net/forum?id=yzkSU5zdwD>.

[10] Including Anthropic's Claude, Microsoft's Copilot, Meta's range of Llama models, and Google's Bard and Gemini.

[11] Since running the study in late 2023, OpenAI have released newer versions of its LLMs, such as GPT-4-Turbo and GPT-4o, which OpenAI claims are more capable and have more recent world knowledge.

[12] Such as Daniel Martin Katz et al, 'Gpt-4 Passes the Bar Exam' [2023] *382 Philosophical Transactions of the Royal Society A (2024)*; Jaromir Savelka et al, 'Explaining Legal Concepts with Augmented Large Language Models (GPT-4)' [2023] *arXiv*.

[13] OpenAI, 'GPT-4 Technical Report' [2023] <https://arxiv.org/abs/2303.08774>.

[14] Such as the technical papers on GPT-3 (Tom Brown et al, 'Language Models Are Few-Shot Learners' (2020) 33 *Advances in neural information processing systems* 1877. and Radford et al (n 8).

[15] Strictly speaking, LLMs process sub-word 'tokens', but that distinction is set aside for the purposes of this article. The first part of the training process is akin to predicting a word hidden by a redaction. For example: 'The capital of [BLANK] is Paris.'

[16] OpenAI (n 13).

stage, human judgements about the quality of the model's output in response to queries are used to train the model further, so that the knowledge it has acquired at original training stage can be delivered in a format that is preferable to a human user, avoids objectionable content, and is aligned with their intent. It is specifically optimized for dialogue with a human user.[17] It is this finetuning for dialogue that makes the GPT-4 model appropriate for use as the model underpinning the chat-interface in the OpenAI product, ChatGPT.

Because GPT-4 is fine-tuned for dialogue, it can be instructed to solve language tasks in plain English instruction. These instructions are known as 'prompts.' Because prompts are the most accessible way to guide these models to produce a desired output, a dedicated field of research into 'prompt engineering' has emerged. This is the process of optimizing the language in a prompt in order to elicit the best possible performance from an LLM for a particular downstream task.[18]

It is this combination of a vast knowledge-base (from its first training stage on massive corpora) and its dialogue interface (from the RLHF stage) that make GPT-4 highly suitable for the production of persuasive legal writing. It is likely to have encoded within itself knowledge about a wide range of legal concepts, and its dialogue interface permits relatively easy extraction of that knowledge into the desired format.

III PERSUASIVE LEGAL WRITING

In this study, GPT-4 is applied to the task of generating 'persuasive legal writing'.[19] For the purposes of this study, persuasive legal writing is defined as ‘*text written in the legal domain for the purposes of persuasion*.’ Persuasive legal writing can take many forms, from a written submission made during a court case, a judgment, inter-partes correspondence, or policy statements written by legal advocacy groups. It is designed to target and persuade its audience, be they a judge, an opposing solicitor, the general public or, as in the present case, a university lecturer. It is often lengthy and dense with references or citations from supporting evidence.

A brief survey of the literature regarding advocacy, argument and persuasive prose was conducted. This survey has revealed several common elements which contribute to effective persuasive writing in the legal domain. These themes, and how they map onto the marking rubric, are discussed below. Existing studies on the performance of LLMs in each of the constituent elements are also explored.

[17] See Xin Zhao et al (n 6).

[18] See Yongchao Zhou et al, ‘Large Language Models Are Human-Level Prompt Engineers’ [2023] The Eleventh International Conference on Learning Representations <https://openreview.net/forum?id=92gvk82DE->.

[19] The term could be used interchangeably with 'written advocacy' or 'legal argumentation'.

A *Structure and Argument*

Persuasive writing must be well structured. ‘Structure is important,’ notes Davies J in a guide to persuasive written advocacy:[20]

> ‘Written work that is dense, impenetrable, lacking cohesion or badly structured will rarely be useful and sometimes may be counter-productive. A valuable opportunity to persuade will have been wasted, sometimes irredeemably.

The ability of an LLM to produce a short essay with an introduction, body paragraphs and a conclusion is a good proxy for its ability to write in a structured manner. As noted in Section IV, below, recent studies indicate that LLMs are capable, if not highly capable, of producing well-structured essay-style writing up to at least several hundred words in length. However, the ability of the models to produce longer form content has not been properly explored.

Persuasive writing must also present an argument. The Stanford Encyclopedia of Philosophy defines argumentation as ‘the communicative activity of producing and exchanging reasons in order to support claims or defend/challenge positions.’[21] Recent studies have shown GPT models are capable of performing various reasoning tasks.[22] These include word problems, typically in the form of a short scenario with multiple choice outputs. Findings from these studies claim that some LLMs have reasoning capabilities, with performance improving with each newer model.

However, a general reasoning ability may not necessarily translate into reasoning ability in the legal domain. Indeed, whether legal reasoning differs fundamentally from 'ordinary' or 'scientific' reasoning, and its processes, are subject to centuries of debate.[23]

Several papers have explored the legal reasoning capability of LLMs. The results have been mixed. Explicit tests of GPT-3's ability to answer logic problems based on synthetic statutes was explored by Blair-Stanek et al.[24] They found that the model outperforms previous benchmarks, but still makes clear errors. As also noted in the following section, several papers have explored the models' performance on law school exams. Good performance on these exams requires a student to analyze a given set of facts, apply their knowledge of the law to those facts, and draw a legal conclusion – the key elements of legal reasoning. Whilst the models are unlikely to follow the same underlying reasoning

[20] Justice Jennifer Davies, ‘Effective and Persuasive Written Advocacy’ (Speech, 7 August 2013).

[21] Stanford Encyclopedia of Philosophy (2021) ‘Argument and Argumentation’.

[22] For example, Hanmeng Liu et al, ‘Evaluating the Logical Reasoning Ability of Chatgpt and Gpt-4’ [2023] *arXiv*.

[23] See Phoebe Ellsworth, ‘Legal Reasoning’ in Keith J Holyoak and Robert G Morrison (eds), *The Cambridge Handbook of Thinking and Reasoning* (Cambridge University Press, 2005) 685.

[24] Andrew Blair-Stanek, Nils Holzenberger and Benjamin Van Durme, ‘Can GPT-3 Perform Statutory Reasoning?’ *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (ACM, 2023) 22 <https://dl.acm.org/doi/10.1145/3594536.3595163>.

process as a human, the studies suggest that the models exhibit a decent ability to mimic these legal reasoning steps, albeit with room for improvement.

B *Knowledge and Understanding*

The factual correctness of written material has a strong bearing on its persuasiveness. Assertions must be factually correct because if they are not, and the reader discovers that they are not, the author's credibility will be undermined. This is likely to affect not only the persuasiveness of the specific point, but the persuasiveness of the essay in its entirety. Former Justice of the High Court of Australia, Kirby J, called credibility an advocate's 'most priceless possession.'[25]

In a study by Savelka et al, an LLM was tasked with explaining how a key term from US statute was used in caselaw.[26] The authors found that, despite responses appearing highly plausible, detailed analysis uncovered limitations in the factual accuracy of the explanations. After testing several chatbots, including ChatGPT, on verifiable questions about random US federal court cases, Dahl et al found that 'legal hallucinations are alarmingly prevalent.' [27] Studies across other disciplines have also found that, when asked to cite sources, LLMs commonly either fabricate sources entirely or conflate multiple sources into a novel hybrid.[28]

Therefore, the tendency of LLMs to make false assertions and to invent sources presents a clear challenge to using LLMs to produce persuasive legal writing.

C *Critical Analysis and Original Reflection*

Excellent persuasive writing requires more than just well-structured, well-reasoned and factually-accurate prose. Aristotle suggested that argument also requires 'pathos', or empathy. Mason J suggests that 'persuasion calls not only for mastery of the materials, but also for an element of constructive imagination and boldness of approach.'[29] In Law and Literature, Mr Justice Cardozo suggests that legal opinions are necessarily persuasive documents and that in order to 'win its way', an opinion must draw upon 'the impressive virtue of

[25] Michael Kirby, 'Rules of Appellate Advocacy: An Australian Perspective' (1999) 1(2) *Journal of Appellate Practice and Process* 227.

[26] Savelka et al (n 12).

[27] Matthew Dahl et al, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' [2024] *arXiv*.

[28] See, eg, Hussam Alkaissi and Samy I McFarlane, 'Artificial Hallucinations in ChatGPT: Implications in Scientific Writing' (2023) 15(2) *Cureus*; David Pride, Matteo Cancellieri and Petr Knoth, 'CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering' [2023] *International Conference on Theory and Practice of Digital Libraries* 146.

[29] AF Mason, 'The Role of Counsel and Appellate Advocacy' (1984) 58(10) *Australian Law Journal* 537.

sincerity and fire, or the mnemonic power of alliteration and antithesis, or the terseness and tang of the proverb and the maxim.'[30]

LLMs have been shown capable of exhibiting some form of empathy, creativity and reflective nous along these lines. Ayers et al posed a series of questions from patients about medical issues on a social media forum and had ChatGPT produce answers to them.[31] The chatbot answers were preferred over those of actual physicians and rated higher in empathy. Haase et al implemented a test requiring participants to generate novel uses for a range of everyday objects.[32] They found that ideas generated by GPT were as creative as any produced by humans, giving doubt to the previously widespread view that AI cannot be creative. Li et al found that GPT-4 can produce reflective writing.[33] Howe et al had ChatGPT-4 produce responses to reader questions from ten newspaper columns by 'agony aunts' like Dear Abby. Comparing same-length responses of the agony aunts to the chatbot, judges rated the system roughly equally empathetic but preferred its answers (even when they were told that some of the answers were not human).[34] These studies show that their LLMs are capable of mimicking the creative, imaginative and empathetic behavior of humans.

IV Recent applications of LLMs to Essay Writing and Law School Exams

A number of studies have examined the ability of an LLM to produce short essays. Yeadon et al explored the capability of a GPT-4 predecessor model to produce short essays for a first-year university subject called 'Physics in Society.'[35] Despite being an earlier model, the essays received first-class grades. Herbold et al produced hundreds of high school essays using a range of OpenAI models and had them scored against non-native English-speaking students.[36] The GPT-4 essays received the highest grades, followed by those from GPT-3.5 (another predecessor model) and, lastly, those produced by the actual students.

Several papers have explored GPT's performance specifically on law school and bar exams. One of the principal papers is 'GPT-4 Passes

[30] Benjamin N Cardozo, 'Law and Literature' (1938) 48 *Yale Law Journal* 489.

[31] John W Ayers et al, 'Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum' (2023) 183(6) *JAMA Internal Medicine* 589.

[32] Jennifer Haase and Paul HP Hanel, 'Artificial Muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity' [2023] *arXiv*.

[33] Yuheng Li et al, 'Can Large Language Models Write Reflectively' (2023) 4 *Computers and Education: Artificial Intelligence* 100140.

[34] Piers Douglas Lionel Howe et al, 'ChatGPT's Advice Is Perceived as Better than That of Professional Advice Columnists' (2023) 14 *Frontiers in Psychology* 1281255.

[35] Will Yeadon et al, 'The Death of the Short-Form Physics Essay in the Coming AI Revolution' (2023) 58(3) *Physics Education* 035027.

[36] Steffen Herbold et al, 'AI, Write an Essay for Me: A Large-Scale Comparison of Human-Written versus ChatGPT-Generated Essays' [2023] *arXiv*.

the Bar Exam' by Katz et al.[37] These findings were touted by OpenAI when it released GPT-4.[38] They claimed 90th percentile performance on the US bar exam. This required the model to answer multiple choice, short answer and longer form open ended questions. All questions require the application of legal reasoning in order to produce a correct answer. Whilst subsequent work has identified challenges in documenting and verifying the 90th percentile claim,[39] it nonetheless suggests the model is very capable of passing difficult legal examinations and exhibits a substantial depth of factual knowledge of US law.

Other studies have also explored GPT's ability by putting it to task on law school exams. These results have not been as glowing as Katz et al, but nonetheless also suggest at least a passable ability of the models to perform legal reasoning. Choi et al studied GPT-4's performance on a spread of law school exams from the University of Minnesota.[40] The model showed only average (C+) performance. Blair-Stanek et al similarly assessed performance on University of Maryland law school exams, showing mixed results but uniformly below average, akin to 'a bright student who never made it to class.'[41] Hargreaves put ChatGPT to task on a wide range on common-law, English-language law school exams from the Faculty of Law of the Chinese University of Hong Kong.[42] He found the answers ranged from strong answers to failing answers across different subjects. He found that whilst the technology is 'incredibly impressive', it can give entirely incorrect answers, invent cases and fail to spot obvious issues. These results, again, suggest an ability to perform legal reasoning, but leave room for improvement.

V METHODOLOGY

A *Producing Long Form Content with LLMs*

Much persuasive legal writing is lengthy. Policy papers, court submissions, court judgments and yes, law school essays, particularly in complex matters, often run well into the thousands of words. However, there is very little literature focused on the ability of models to produce long form content.

For the purposes of this paper, 'long-form' content is considered to be in excess of approximately 750 words. Papers which have explored

[37] Katz et al (n 12).

[38] 'GPT-4'. *OpenAI* (Web Page, 14 March 2023) <https://openai.com/research/gpt-4>.

[39] Eric Martínez, 'Re-Evaluating GPT-4's Bar Exam Performance' [2023] *Artificial Intelligence and Law (2024)*.

[40] Jonathan H Choi et al, 'ChatGPT Goes to Law School' (2021) 71(3) *Journal of Legal Education* 387.

[41] Andrew Blair-Stanek et al, 'GPT-4's Law School Grades: Con Law C, Crim C-, Law Econ C, Partnership Tax B, Property B-, Tax B' [2023] *SSRN Electronic Journal.*

[42] Stuart Hargreaves, '"Words Are Flowing Out Like Endless Rain into a Paper Cup": ChatGPT & Law School Assessments' (2023) 33(1) *Legal Education Review* 69.

the ability of LLMs to generate essay-style responses greater than about 600 words have not been identified.[43]

There are a number of factors which limit the ability of LLMs to produce high-quality longer content. The first is the model 'context window.' This refers to the total size of the input sequence the model can process. The LLM will not be able to 'see' anything outside of this window, and therefore is unable to produce new content which is consistent with it.[44] Whilst the latest models have much larger context windows, LLMs are not necessarily able to maintain optimum reasoning performance over the entire window. Liu et al showed that some models show a notable drop in reasoning ability and performance as prompt size grows, and that reasoning ability drops in the 'middle' of the prompt window.[45] Finally, it is practically difficult to prompt certain models to generate lengthy content in a single inference.[46] Preliminary testing for this study suggests that, for GPT-4, this is typically in the order of a maximum of 750 words, notwithstanding explicit instructions to the contrary in the prompt.

All LLMs, including GPT-4, therefore have an upper limit on the size of the text that it can produce and reason over in a single inference whilst maintaining optimum performance. To produce longer, high quality content, it is necessary to develop techniques which combine content produced in multiple inferences.

A method to produce long-form content was developed for this study. The methodology has three steps. It is inspired by how university students are encouraged to write good essays — namely, to produce a high-level outline of the arguments to be made before writing the content in detail:[47]

1. **Outline:** The model is first prompted to produce a high-level outline of the essay based on the essay topic, the adopted persona of the model and a short selection of references which may be drawn upon in the response. An example of the prompt used to generate this outline is shown in Figure 1.

[43] Papers looking at the ability of the models to produce essay-style content use models which produce text in the range of about 300 words (such as Yeadon et al (n 35)) to about 600 words (such as Katz et al (n 12)).

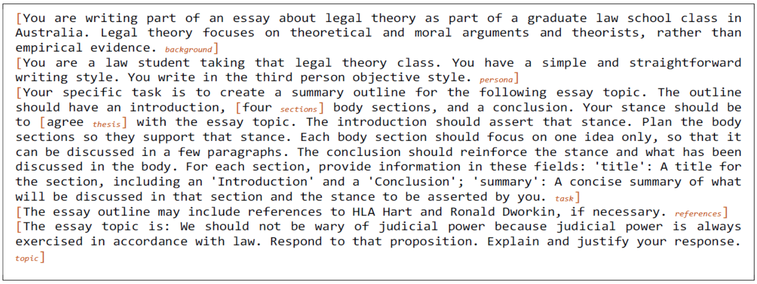
[44] The context window varies amongst LLMs. GPT-2, for example, had a context window of 1,024 tokens (approximately 750 words), whilst the latest OpenAI model at the time of writing, GPT-4o, has a context window of 128,000 tokens (approximately 96,000 words).

[45] Nelson F Liu et al, 'Lost in the Middle: How Language Models Use Long Contexts' [2023] *arXiv*.

[46] This study defines a 'single inference' as the generation of a single block of new text by the LLM (as opposed to repeated prompting or back-and-forth chat with a product such as ChatGPT).

[47] For example, Marcus Cleaver, 'How to Write a First Class Law Essay', *UK Law Weekly* <https://uklawweekly.com/wp-content/uploads/2020/11/How-To-Write-First-Class-Law-Essays.pdf>.

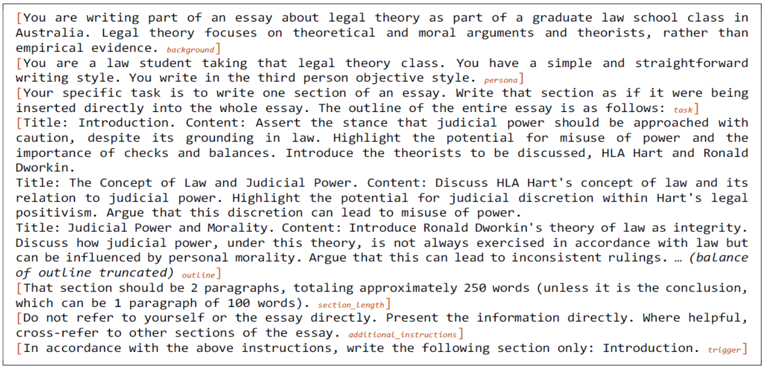
Figure 1
Prompt used to generate an outline for the essay topic on 'judicial power'


[You are writing part of an essay about legal theory as part of a graduate law school class in Australia. Legal theory focuses on theoretical and moral arguments and theorists, rather than empirical evidence. background]
[You are a law student taking that legal theory class. You have a simple and straightforward writing style. You write in the third person objective style. persona]
[Your specific task is to create a summary outline for the following essay topic. The outline should have an introduction, [four sections] body sections, and a conclusion. Your stance should be to [agree thesis] with the essay topic. The introduction should assert that stance. Plan the body sections so they support that stance. Each body section should focus on one idea only, so that it can be discussed in a few paragraphs. The conclusion should reinforce the stance and what has been discussed in the body. For each section, provide information in these fields: 'title': A title for the section, including an 'Introduction' and a 'Conclusion'; 'summary': A concise summary of what will be discussed in that section and the stance to be asserted by you. task]
[The essay outline may include references to HLA Hart and Ronald Dworkin, if necessary. references]
[The essay topic is: We should not be wary of judicial power because judicial power is always exercised in accordance with law. Respond to that proposition. Explain and justify your response. topic]

The 'fields' that were defined to populate the prompt are labelled in sub-script.

2. **Content:** Each paragraph of the essay is then produced independently, based upon the summary of that paragraph generated in step 1. The prompt includes the entire outline from step 1, instructions on persona of the model, instructions on which section is to be written, and details such as the word count. An example content prompt is shown in Figure 2.

Figure 2
The common prompt used to generate the first section of the essay on 'judicial power'


[You are writing part of an essay about legal theory as part of a graduate law school class in Australia. Legal theory focuses on theoretical and moral arguments and theorists, rather than empirical evidence. background]
[You are a law student taking that legal theory class. You have a simple and straightforward writing style. You write in the third person objective style. persona]
[Your specific task is to write one section of an essay. Write that section as if it were being inserted directly into the whole essay. The outline of the entire essay is as follows: task]
[Title: Introduction. Content: Assert the stance that judicial power should be approached with caution, despite its grounding in law. Highlight the potential for misuse of power and the importance of checks and balances. Introduce the theorists to be discussed, HLA Hart and Ronald Dworkin.
Title: The Concept of Law and Judicial Power. Content: Discuss HLA Hart's concept of law and its relation to judicial power. Highlight the potential for judicial discretion within Hart's legal positivism. Argue that this discretion can lead to misuse of power.
Title: Judicial Power and Morality. Content: Introduce Ronald Dworkin's theory of law as integrity. Discuss how judicial power, under this theory, is not always exercised in accordance with law but can be influenced by personal morality. Argue that this can lead to inconsistent rulings. … *(balance of outline truncated)* outline]
[That section should be 2 paragraphs, totaling approximately 250 words (unless it is the conclusion, which can be 1 paragraph of 100 words). section_length]
[Do not refer to yourself or the essay directly. Present the information directly. Where helpful, cross-refer to other sections of the essay. additional_instructions]
[In accordance with the above instructions, write the following section only: Introduction. trigger]

The 'outline' section (truncated in the image) is the verbatim output of the Outline produced in step 1.

3. **Concatenation:** Each separate section generated above is then concatenated into the final product.

The prompts referred to above were developed by the authors. This process sought to mimic that which a bright but non-committed student might undertake when using a tool such as ChatGPT to assist them with an assessment on a time-constrained task. Determining what that process looked like was necessarily more art than science. There are countless ways in which a university student may apply the technology to assist them in writing an essay, including differing combinations of prompts, fact checking the model's outputs, brainstorming ideas or essay plans with the model, and having it rewrite or 'polish' writing originally prepared by the student. This combination will depend on the student's time availability, commitment and understanding of the technology. Unfortunately, there was currently little empirical work on the exact techniques and prompt styles used by students. It was assumed that the time available for prompt experimentation was limited (as a student will be in exam conditions, and a busy student is likely to be even outside of exam conditions), but that the LLM user had some experience with the tools and an awareness of the LLMs ability and shortcomings (as bright, young students may do.)

An iterative prompt engineering exercise was then conducted in accordance with these assumptions. The quality of the LLM output for each set of prompts was evaluated, and the prompts were iteratively improved. Evaluation of the interim output was conducted with reference to the three components of persuasive legal writing identified above. Trends in the effectiveness of the different prompt styles were noted. Google searches were used to spot check the factual accuracy of some citations and sourcing produced by the LLM, although not every claim was fact-checked. Further observations on this prompt engineering phase are made in the Discussion section, below.

B *Long Form Content Variant*

A second slight variant of the method described above was also developed. This amended method was used to produce the second essay for each essay topic. Both methods follow the same 'Outline' process from Step 1, above. It differed only in that the text produced by the LLM for earlier paragraphs of the essay remained 'visible' to the LLM in the prompt when it produced later paragraphs.[48] For example, the content produced for an 'introduction' section of an essay remained 'visible' to the LLM in the prompt when it was asked to produce the subsequent 'body paragraph' of the essay. The intuition behind this variant was that each subsequent paragraph may be more coherent with previous sections if those previously produced sections are 'visible' by the LLM, rather having each paragraph or section produced in isolation and then concatenated.

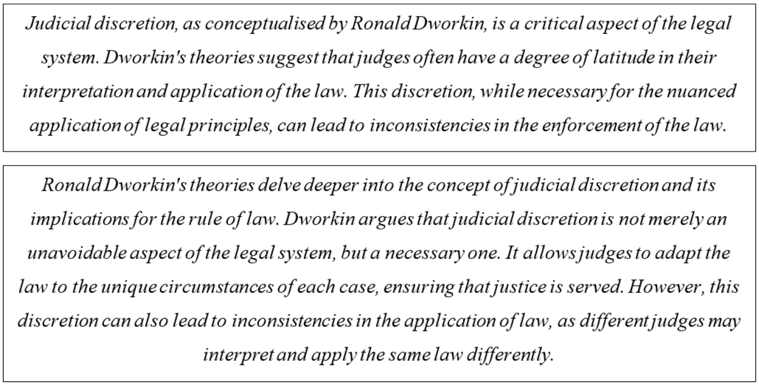
An example of how these slight difference in the prompt structure can result in visibly different text output is shown in Figure 3. Users of a product such as ChatGPT are likely to observe similar slight differences in output even when they repeatedly enter the *exact* same prompt into the chat interface. This is due to the LLM's non-determinism and the ChatGPT temperature settings.[49] (No significant

[48] 'Visible' in this context means that the text is contained in a prompt, the length of which remains within the LLM's 'context window'. In the case of essays of less than 1,500 words, this was well within GPT-4's context window of 8,192 tokens (approximately 6,000 words).

[49] 'Non-determinism' describes the phenomenon where an LLM randomly returns different outputs for the same input prompt. See Shuyin Ouyang et al, ‘LLM Is Like a Box of Chocolates: The Non-Determinism of ChatGPT in Code Generation’ [2023] *arXiv*. The 'temperature' of a GPT model is a user setting which controls how random

difference in the grades of the output from these methods was subsequently found, and a much larger sample size and dedicated study may be necessary to tease out any grading differences in these variant methods.)

Figure 3
Slight variations in LLMs prompts can lead to a different output



Judicial discretion, as conceptualised by Ronald Dworkin, is a critical aspect of the legal system. Dworkin's theories suggest that judges often have a degree of latitude in their interpretation and application of the law. This discretion, while necessary for the nuanced application of legal principles, can lead to inconsistencies in the enforcement of the law.

Ronald Dworkin's theories delve deeper into the concept of judicial discretion and its implications for the rule of law. Dworkin argues that judicial discretion is not merely an unavoidable aspect of the legal system, but a necessary one. It allows judges to adapt the law to the unique circumstances of each case, ensuring that justice is served. However, this discretion can also lead to inconsistencies in the application of law, as different judges may interpret and apply the same law differently.

Figure 3 shows a paragraph from the body section of one of the GPT-4 produced essays is shown, the first textbox using the main method, and second textbox using the slight variant.

C *Evaluating the GPT-4 Essays*

GPT-4's ability to produce long form persuasive writing was evaluated by having it generate an essay for a 'Legal Theory' examination. Legal Theory is a postgraduate-level unit at the University of Melbourne.[50] The subject explores ways to think about legal concepts, institutions, processes, roles and values in the law. The essay questions are open ended, invite students to argue a position and are typically one or two sentences in length. Whilst there is no minimum word count, high achieving students typically produce essays in the range of 1,000–1,500 words. Any citations must be clear and consistent. Students are given three hours to produce these essays under exam conditions.

The exam-based essay format was chosen in lieu of alternative assessment formats for several reasons. First, the performance of GPT models on multiple choice and short answer questions had already been examined by others (see above Section IV) and would not have provided as novel a contribution. Second, an essay format allowed us to test complex legal reasoning on open ended questions. Finally,

the token outputs are. Higher temperature results in more diverse outputs each time an inference is made. Whilst the temperature can be controlled when GPT-4 is accessed via an API, a user of ChatGPT (at the time of publication) has no control over this setting. These factors combined mean that users of ChatGPT will likely receive slightly different responses each time an *identical* prompt is repeated.

[50] The University of Melbourne (n 5).

students writing an *exam*-based essay have less time to produce their essay and elaborate citations compared to, say, a take-home research essay. As such, a take-home research essay may have set the bar too high for the current generation of LLMs. Of course, further studies exploring an LLM's competency in these other assessment formats would help produce a richer benchmarking of its abilities, as would a comparison against similar assessments at an undergraduate level.

Four Legal Theory essays were produced with GPT-4. Two exam questions were used. Two essays were produced for each question.[51] Each essay was assessed independently by two different graders.

The exam questions were:

1. 'We should not be wary of judicial power because judicial power is always exercised in accordance with law. Respond to that proposition. Explain and justify your response.'
2. 'Since the law is not morally-neutral, legal education must necessarily include an engagement with morality. Do you agree? Explain and justify your response.'

D *Student Benchmarks*

The GPT-4 essays were benchmarked against essays written by actual students who had previously undertaken the Legal Theory course. Four essays were selected from two second-class honors students (an H2A and an H2B student) who had written on the selected topics.[52] These were not the best performing students in the class. The H2A student was nonetheless above average, and the H2B student in the middle of their cohort.[53] The essays were anonymized. They averaged 1,185 words each. Minor formatting changes were made to the essays to ensure consistent formatting both amongst the students and the artificially generated essays.

The handling of citations and footnotes proved challenging. On the one hand, the student essays contained dense, accurate, pinpointed footnoting. On the other, the testing during the development of the prompts showed that GPT-4 had difficulty producing content which could be reliably cited. This is a well-known shortcoming of LLMs.[54] This was the case whether the citation was 'in text' or contained in a footnote.

Ultimately the footnotes were maintained in the student essays, because the graders needed to know if the writer was proposing an original idea, and whether the writer was relying on valid evidence. The

[51] For each topic, one essay was produced using the main method, and a second essay was produced using the variant method, as described in sub-sections A and B, above.

[52] At the University of Melbourne, H2A means 'Second Class Honours Division A' (75% - 79%). H2B means 'Second Class Honours Division B' (70% - 74%).

[53] Due to resourcing restraints the study was limited to a comparison with average and above-average performing (but not top-of-the-class) students. Further studies against a more diverse competency range would be welcome and help produce a richer benchmark of the technology.

[54] See, eg, Alkaissi and McFarlane (n 26); Pride, Cancellieri and Knoth (n 28).

deletion or amendment of footnotes could affect the substantive quality of the student essay and its grading.

GPT-4 could similarly have been prompted to include citations. However, it was decided that any inaccuracies would likely have been noticed by a human grader familiar with the course material, and that the discovery of inaccurate or non-existent sources would have a more detrimental effect on the essay grade than the mere absence of citations. Therefore, the prompts used in this study did not ask GPT-4 to produce footnotes. The result was an obvious visual discrepancy between the students' essays and the GPT-4 essays, in that the former contained footnotes on each page, and the latter did not. Further comments on this are made in the Results section, below.

E *Essay Grading*

Four graders participated in the study. The graders were staff at the University of Melbourne who had prior experience grading Legal Theory exams during the course using the same grading rubric. The graders were informed that their evaluations were being used for comparison with *evaluations* produced using a computational text analysis tool. They were not told that any of the essays they were evaluating were produced artificially. They were not told about the provenance of any of the essays, either human, LLM, or otherwise.

Each grader marked two essays on each topic - one artificially generated and one student essay. Each grader marked four essays in total. The essays were distributed so that each artificially generated topic-method combination was independently evaluated against both the H2A student and the H2B student essay.

The graders were instructed to:

1. Read each essay;
2. Rate the essay either Excellent, Very Good, Good, Satisfactory or Needs Improvement, in each category of the following marking rubric:
    a. Argument and structure: ‘Ability to develop an organized and reasoned response to the selected topic, justified by reference to relevant theorists and theoretical approaches.’ (*Argument and Structure*)
    b. Knowledge and understanding: ‘Knowledge and understanding of theoretical texts and arguments studied in the course relevant to the selected question.’ (*Knowledge and Understanding*)
    c. Critical analysis and original reflection: ‘Ability to critically analyze, compare, evaluate, situate and comment on theoretical arguments and accounts of law.’ (*Critical Analysis and Reflection*)
3. Provide an overall grade for the essay of either H1, H2A, H2B, H3 or Fail.

The graders could also, optionally, provide additional comments on the essay.

F *Evaluation Metrics*

The individual criteria and overall grade were converted to numerals. The criteria 'Excellent', 'Very Good', 'Good', 'Satisfactory', or 'Needs Improvement' were converted to 5, 4, 3, 2 and 1, respectively. The overall grade of H1 (excellent), H2A (very good), H2B (good), H3 (competent) or Fail were converted to 5, 4, 3, 2 and 1, respectively. A qualitative analysis of the comments was performed by manual review and comparison. A sentiment analysis on the comment text was also performed. Sentiment analysis is the task of computationally categorizing the writer's attitude in a piece of text, typically into a 'positive', 'neutral' or 'negative' classification.[55] The VADER automated sentiment analysis tool was used.[56] VADER produces a single measure of the sentiment of the text, normalized from -1.0 (very negative) to +1.0 (very positive). The VADER sentiment scores were not manually validated.

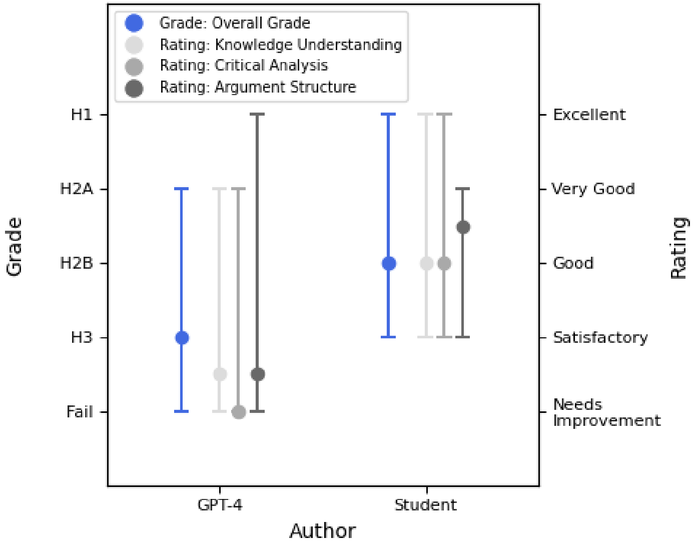
VI RESULTS

Sixteen graded essays were received from four graders. The median Grade for the GPT-4 essays was H3 (competent). The median Grade received for the student essays in this study was H2B (good). The minimum, median and maximum values for the GPT-4 and Student essays are shown in Figure 4.

[55] See Venkateswarlu Bonta, Nandhini Kumaresh and N Janardhan, 'A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis' (2019) 8(S2) *Asian Journal of Computer Science and Technology* 1.

[56] VADER is a simple, yet popular, rule-based automated model for sentiment analysis. It uses a predefined vocabulary and heuristics to assign sentiment scores to a string of text (for example, the string 'great!' may receive a positive score because it contains the word 'great', followed by an exclamation.) See C Hutto and Eric Gilbert, 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text' (2014) 8(1) *Proceedings of the International AAAI Conference on Web and Social Media* 216 ('VADER'). VADER has been used in previous studies including tools which auto-evaluate essay writing. See Harneet Kaur Janda et al, 'Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation' (2019) 7 *IEEE Access 108486* .

Figure 4
Minimum and maximum (denoted with '-') and median (denoted by 'o') values by author[57]



The student essays median grade (H2B) was one clear grade category higher than the GPT-4 essays (H3). The results also show slightly elevated ratings for Argument and Structure for the GPT-4 essays compared to the other criteria, albeit marginal.

Optional comments were received from the graders for 14 of the 16 essays, six of which were comments for the GPT-4 essays and eight of which were for student essays.

The eight comments for the student essays were mixed, but generally constructive. The H2A student essays were mostly praised, save for a suggestion on signposting more and one comment which critiqued the overuse of examples without making a 'substantive point.' The H2B student essays were complimented on their structure and argument, but it was noted that they had some stylistic errors and poor editing. One comment noted the H2B essay was under-argued and had problems with relevance.

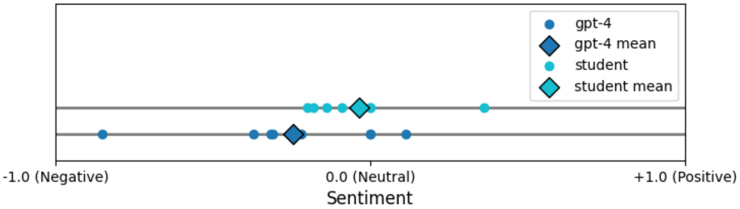
The six comments for the GPT-4 essays were generally negative. They noted the essays had 'thin understanding', 'very little here', and showed a 'lack of theoretical depth.' As expected, they highlight the absence of citations (i.e. the absence of footnotes, as noted above). It appears likely from these comments that the absence of footnotes negatively affected the GPT-4 essay grades. The graders may have seen the absence of citations as an indication of a lack of effort, or even the suspected use of ChatGPT or similar tools (although it should again be noted that the graders were not explicitly told about the provenance of

[57] Some grader feedback was manually adjusted to the nearest fixed grading class to match the grader instructions, as follows: * 'Pass/H3' converted to H3; $^{+}$ 'Fail/NI' converted to Needs Improvement; $^{\#}$ 'Pass' converted to H3

the essays.) However, there was no control group to measure the size of any such effect. The limited positive feedback for the GPT-4 essays was that one of them was 'clearly written', and another was 'well argued.'

A sentiment analysis was conducted on the comments. The comment sentiment means are plotted in Figure 5. The mean for the GPT-4 comments is -0.245, lower (i.e. more negative) than the mean of the student comments at -0.037.

Figure 5
Average sentiment of grader comments



VII Discussion

A *Ability of LLMs to Generate Long Form Persuasive Legal Writing*

The results are consistent with existing literature from Choi[58], Blair-Stanek[59] and Hargreaves[60] who in the academic legal context found the models showed variable performance which was passable, but not excellent.

The finding that the essays were (slightly) more proficient at Argument and Structure than at the other areas is to be expected, given earlier studies confirming the model's ability to produce structured argument. Similarly, it was no surprise that the model did not excel at Knowledge and Understanding, given the well-known inability of LLMs to consistently produce factually accurate content. One illustrative comment noted that a GPT-4 essay *"gets Dworkin badly wrong."*

As for Critical Analysis and Reflection, there was no suggestion of the flashes of brilliance, creativity or originality from GPT-4 that studies in other fields suggest might be expected. This muted performance may have been the result of the decisions made during the prompt engineering phase, discussed below, that sought to steer the output of the models so that it was focused on the curriculum. On the one hand this reduced the scope for the model to generate random or off-topic content, but on the other may have muted the model's 'flair.'

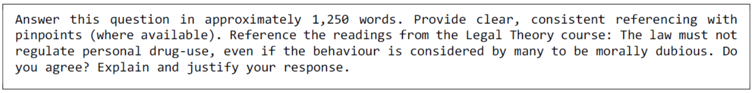
[58] Choi et al (n 40).
[59] Blair-Stanek et al (n 41).
[60] Hargreaves (n 42).

B *Prompt Engineering*

The final prompts which were fed to GPT-4 to generate the essays were the result of a prompt engineering phase of iterated stepwise improvements and variations of prompts. A competent student familiar with the technology, using an LLM via a chatbot interface such as ChatGPT, is likely to also undertake a similar iterative process. A number of observations on this process are made below.

The prompt engineering phase began by experimenting with prompts that mirror the simple instructions given to students undertaking the course. An example is shown in Figure 6. The prompt is short and presumes knowledge about the context of the task which GPT-4 is unlikely to have (such as knowledge of the 'Legal Theory' course content).

Figure 6
Example of an early prompt whilst developing the prompt templates



```
Answer this question in approximately 1,250 words. Provide clear, consistent referencing with
pinpoints (where available). Reference the readings from the Legal Theory course: The law must not
regulate personal drug-use, even if the behaviour is considered by many to be morally dubious. Do
you agree? Explain and justify your response.
```

The outputs from these simple, short prompts were well structured and well written. They did address the essay topic and make logical arguments. However, they had the following shortcomings:

1. They were short. Despite many variations on the word count instructions, it was difficult to have GPT-4 produce content in a single inference that exceeded about 750 words;
2. They were often US-centric. For instance, they would cite legal cases such as ‘Roe v Wade’ and ‘Brown v Board of Education’.[61] Whilst drawing parallels to these cases might be a creative and interesting take on an essay topic, it was unusual in an Australian context. (This tendency could be seen as a form of geographic-bias in the model, as discussed in greater detail in the following section.);
3. They included material that was outside of the content matter of the Legal Theory course. For instance, the prompt relating to the regulation of drug-use would make reference to empirical evidence on the Portuguese decriminalization of drug-use, which was not only not part of the curriculum, but also an empirical matter somewhat out of place in an essay on legal theory. To the extent they included theoretical content, it was often theorists which were not covered in the course curriculum, such as Thomas Aquinas.

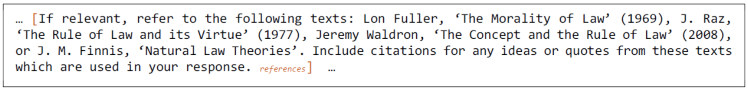
[61] Dahl et al (n 27), suggests that the bias of GPT-3.5, the precursor model to GPT-4 used in this study, may be skewed towards the most well-known decisions of the American legal system.

Further experiments were conducted to build up the prompt templates to further 'steer' the output. Explicit instructions were added about the essay being written for a course in Australia and requiring a focus on theoretical arguments rather than empiricism. The topic headings that were studied in the course curriculum were listed, and the model was asked to produce content which was relevant to those topics.

GPT-4 was responsive to these prompt changes, in that it focused on theory and mentioned the curriculum topics in its output. However, the output did not suggest any more than a superficial understanding of the course material. In contrast, the argument and citations in the honours student essays were much more detailed. They contained pinpointed references to academic papers, quotations relevant to the assertions in the essay body, and showed a command of the course material.

Further 'steering' the GPT-4 model to refer specifically to theorists, papers and quotations from the Legal Theory reading guide were tried. An example of the experiment with further steering of the output is shown in Figure 7. This prompt included references to the specific content from the subject reading guide for the chapter relevant to the essay topic.

Figure 7
An example 'references' component of the prompt template used during the prompt engineering phase



… [If relevant, refer to the following texts: Lon Fuller, ‘The Morality of Law’ (1969), J. Raz, ‘The Rule of Law and its Virtue’ (1977), Jeremy Waldron, ‘The Concept and the Rule of Law’ (2008), or J. M. Finnis, ‘Natural Law Theories’. Include citations for any ideas or quotes from these texts which are used in your response. *references*] …

It was found that being overly specific (such as in this example) led to the output which focused too heavily on the prompted content, and often hallucinated.

The further 'steering' of the model output proved challenging. When prompts that were very specific (such as the example in Figure *7*) were used, it was found that the model output focused too heavily on the prompted content. The output forcibly shoe-horned the references into the arguments, irrespective of whether they were relevant or not, and often hallucinated content which it claimed to be from the sources that were included in the prompt. It was evident by simple Google searches of the source material that many of the quotations cited did not exist in the original documents. From time to time, when steered in this manner, the model also entirely mischaracterized a theorist's point of view, confidently asserting the complete opposite position of the theorist.

The challenge, therefore, was to include enough light steering to guide the output towards the desired themes, ideas and source material, but would not overdo guidance. If the prompt was too generic, the output may be well written and even creative, but could stray too far from the desired form and the substance required by the curriculum. But if the prompt was overly prescriptive of the desired form and content, it appeared to over-compensate in the direction of the request, often hallucinate, and deaden the model's creativity.[62]

For the final essay productions, a balanced prompt was included that mentioned a few of the key theorists which were focused on in the course, but did not specify any works or ideas of theirs. Doing so appeared to allow GPT-4 to draw upon its own knowledge of that author's work and integrate it into the essay thesis in a more natural way.

The ordering of the prompt components also made a difference. When the prompt was lengthy, the model sometimes ignored certain instructions which were buried in the middle of the prompt window. However, the model did comply with the instruction if it was reiterated at the end of the prompt window.[63]

Finally, it is worth noting that the number of variations of possible prompts for a model such as GPT-4, and possible generated outputs, is practically limitless.[64] It is therefore very difficult given the current understanding of LLMs to make any definitive claims about the absolute performance (or lack thereof) of certain LLMs or prompt templates. Indeed, the construction of scientific approaches to prompt engineering and output evaluation is an active research topic.[65]

C *Bias in LLMs and its impact on legal scholarship*

It was not the goal of this study to examine biased outputs from LLMs. Nonetheless, our observations, discussed below, act as a helpful segue into an important discussion of the implications of LLM-bias on legal scholarship and education.

Bias in computer systems is the systematic and unfair discrimination against certain individuals or groups in favour of others.[66] An LLM can produce text which manifests biases in several ways, such as the association of gendered pronouns or racial descriptions with certain

[62] The tendency of models to produce output which matches a user's belief in preference to ground truth has been referred to as a form of sycophancy. (See Mrinank Sharma et al, ‘Towards Understanding Sycophancy in Language Models’ [2023] *arXiv*).

[63] This is consistent with Liu et al (n 45).

[64] This is the result of the vast number of ways the model's large vocabulary can be arranged within the model's large context window. The resulting 'combinatorial explosion' results in a practically infinite number of prompt combinations.

[65] See, eg, Chirag Shah, ‘From Prompt Engineering to Prompt Science With Human in the Loop’ <https://arxiv.org/abs/2401.04122>.

[66] Whilst the definition of 'bias' can vary across disciplines, this definition is commonly used in computer science literature. It was proposed by Batya Friedman and Helen Nissenbaum, ‘Bias in Computer Systems’ (1996) 14(3) *ACM Transactions on information systems (TOIS)* 330.

traits.[67] Its existence is openly acknowledged by LLM developers.[68] It is a well-studied area in computer science literature. The literature spans applications across many domains including journalism, politics, medicine and, relevantly to this study, education.[69] However, whilst there has been some work examining the biases in earlier AI systems in the legal domain, there are few studies on biases in legal applications specifically with LLMs.[70] Biases in LLM outputs may be reflective of the biases in their training data. This includes historical bias in the data (for example the fact that, historically, Fortune 500 CEO's have overwhelmingly been men), or representational bias in the data (such as a lack of geographically, culturally or linguistically diverse datasets to be used for training).[71]

As noted above, a form of geographic bias in the LLM was observed in our study.[72] Specifically, when citing references in the essays, the LLM had a tendency to cite US caselaw when it had not been asked to do so, and despite there being no indication that the user was US-based.

[67] Yogarajan et al gives an example of a 'continuation' (i.e. a completion of a sentence) by an earlier LLM, GPT-2, to an initial prompt about ‘two men’. When the men were specified as 'brown Maori', the model completed a sentence relating to the conduct of a crime. When the men were specified as 'white kiwi', the model completed the sentence by describing a list of positive attributes: Vithya Yogarajan, Gillian Dobbie and Henry Gouk, ‘Effectiveness of Debiasing Techniques: An Indigenous Qualitative Analysis’ [2023] *arXiv*.

[68] Including OpenAI, for example. See ‘GPT-4 System Card’ <https://cdn.openai.com/papers/gpt-4-system-card.pdf>.

[69] Fang et al identified gender and racial biases when an LLM was prompted to generate artificial news content based on a headline: Xiao Fang et al, ‘Bias of AI-Generated Content: An Examination of News Produced by Large Language Models’ (2024) 14(1) *Scientific Reports* 5224. Rutinowski et al posed questions from the political compass test to ChatGPT, revealing a bias towards progressive and libertarian views: Jérôme Rutinowski et al, ‘The Self-Perception and Political Biases of ChatGPT’ (2024) 2024(1) *Human Behavior and Emerging Technologies* 7115633. Zack et al prompted GPT-4 on various clinical and medical education applications, finding that it failed to appropriately model the demographic diversity of medical conditions, stereotyping demographic presentations: Travis Zack et al, ‘Assessing the Potential of GPT-4 to Perpetuate Racial and Gender Biases in Health Care: A Model Evaluation Study’ (2024) 6(1) *The Lancet Digital Health* 12. In a broad survey on their use in educational applications, Caines et al discuss the use of LLMs in educational content creation, assessment and feedback. They note that even the 'multilingual' LLMs have a strong bias towards English: Andrew Caines et al, ‘On the Application of Large Language Models for Language Teaching and Assessment Technology’ [2023] *arXiv*.

[70] Earlier legal AI systems included the COMPASS recidivism prediction tool, which was deployed in court systems across the US in the 2010s. An analysis by Angwin et al indicated that its predictions were racially biased: Julia Angwin et al, ‘Machine Bias’ (23 May 2016) *ProPublica*; Julia Dressel and Hany Farid, ‘The Accuracy, Fairness, and Limits of Predicting Recidivism’ (2018) 4(1) *Science advances* 5580. In contrast, one of the few studies examining LLM bias in a legal context, a recent preprint from Schwartz et al, suggests that GPT-4 exhibits *no* significant race or gender bias when tasked with hypothetical conviction predictions: Talia Schwartz and Chen Wang, ‘Impartial or Biased? The Effect of Race, Gender, and Priming on AI’s Conviction Predictions’ [2024] *Gender and Priming on Large Language Models’ Conviction Predication (March 31, 2024)*.

[71] See Ninareh Mehrabi et al, ‘A Survey on Bias and Fairness in Machine Learning’ (2021) 54(6) *ACM computing surveys (CSUR)* 1.

[72] There is some emerging literature exploring this concept. See, for example, Rohin Manvi et al, ‘Large Language Models Are Geographically Biased’ [2024] *arXiv*.

This accords with other studies which have suggested a disparity in LLM performance *across* jurisdictional boundaries.[73] It also aligns with earlier work on (non-LLM) AI models used in education which exhibit a US-centric approach.[74] An interesting, if somewhat concerning idea, emerges from this literature. Dahl et al, drawing on the work of Kleinberg et al on algorithmic monocultures, suggests that these LLM biases may instantiate a form of 'legal monoculture'.[75] They state that

> instead of accurately restating the full variation of the law, LLMs may simply regurgitate information from a few prominent members of the response set that they have been trained on, flattening legal nuance and producing a falsely homogenous sense of the legal landscape.[76]

Caines et al considered the educational applications of LLMs in three categories – content creation, assessment, and feedback.[77] Although not observed in this study, it is also not difficult to envisage other LLM biases *potentially* impacting legal education and scholarship in these categories. For example, on content creation, an LLM could conceivably be used by a criminal law teacher to produce hypothetical case studies for her class, into which potential racial stereotypes may be injected. An LLM's hyper-proficiency in English language tasks may incentivize young academics to avoid teaching or examining in low resource languages (such as an indigenous language) where the LLM may be less reliably able to assist with automated-grading.[78] When providing automated feedback, an LLM may unfairly critique legal concepts or research ideas of which it has little knowledge, because they are relatively obscure concepts or derived from predominantly oral

[73] Dahl et al noted jurisdictional differences across US states of hallucination rates on legal case retrieval tasks: Dahl et al (n 27).

[74] Baker et al, reviewing algorithmic bias in education (albeit in 2019, prior to the release of ChatGPT and the recent explosion in interest in such topics), noted an *'intense American focus of research'* as a reflection of where the research was conducted: Ryan S Baker and Aaron Hawn, 'Algorithmic Bias in Education' [2022] *International Journal of Artificial Intelligence in Education* 1.

[75] Dahl et al (n 27), drawing on Jon Kleinberg and Manish Raghavan, 'Algorithmic Monoculture and Social Welfare' (2021) 118(22) *Proceedings of the National Academy of Sciences* e2018340118.

[76] Dahl et al (n 27) 5. This notion seems less fanciful when one considers that there currently exists only a handful of cutting-edge, powerful LLMs. These are produced by large corporations such as OpenAI, Google and Meta. Other specifically legal chatbot products used by law firms and academics, such as Lexis Nexus' Lexis+ AI product, integrate these foundational models into their products: Lexis Nexis, 'LexisNexis Launches Lexis+ AI' (25 October 2023) <https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-launches-lexis-ai-a-generative-ai-solution-with-hallucination-free-linked-legal-citations>. The technical expertise, resources and data required to develop these models impose significant hurdles to new entrants. Whilst the developers are mostly US-based, their user base of students, academics, lawyers and jurists are in diverse, global legal landscapes.

[77] Caines et al (n 69).

[78] Adelani et al found that GPT-4 performs certain labelling tasks worse in low-resource languages (namely, indigenous languages from Brazil and Africa), than in English: David Ifeoluwa Adelani et al, 'Comparing LLM Prompting with Cross-Lingual Transfer Performance on Indigenous and Low-Resource Brazilian Languages' [2024] *arXiv*.

traditions of which little text has historically been produced for inclusion in LLM training data.

Despite an awareness that such biases exist and could impact legal education, the study of the extent and impact of biases in LLMs is more challenging than for earlier systems. LLMs have higher dimensionality than other AI systems, such as those that produce binary outputs or classifications (say, 'low risk', or 'high risk' of recidivism). There is also often no single 'correct' response to an open-ended text question, such as an essay, and evaluating correctness is itself subjective. Further, behaviour that is undesirable in one context can be beneficial in another. For instance, a user may want the LLM to produce responses which are tailored to the background and pedagogical needs of an individual student, or produce creative writing in which hallucinations enrich the output.

There is no easy solution for legal educators. Bias in LLMs derives, ultimately, from training data, architecture and training techniques used by the model developers. As this process is resource intensive and, certainly in the case of OpenAI's frontier models, kept predominantly out of public view, there is little individuals can do to directly rectify its root-causes. Nonetheless, bias across well-studied dimensions of gender and race appear to be reduce in each newer model from the large developers.[79] Further study on LLM bias in the legal and educational domain would be welcome, exploring new dimensions along which biases can manifest themselves in different scholarly use cases. Given new models are periodically released, and each model exhibits different characteristics, studies may need to be reproduced to ensure findings hold across model versions over time. As a basic first step, an awareness of LLM's tendency to produce biased outputs should help legal educators navigate their adoption.

D *Implications for the Legal Profession and Legal Education*

This study should dampen any concerns about LLMs being immediately available to students to produce excellent, H1 quality, long-form persuasive legal writing. This study showed that generating long form content using GPT-4 is difficult using simple prompts. A more complex, multi-step approach is required to produce content at length. This work suggests that whilst concerns such as factual inaccuracies and hallucination may be quelled by tinkering with prompts, this also appears to subdue the creativity and breadth of the content that is produced. Even after these tweaks, the essay cannot be guaranteed to be factually accurate.

We found that producing a decent output is challenging and requires a significant amount of manual tinkering and verification. Unless a powerful and reliably effective prompt template is discovered, a student

[79] See, eg, Anthropic's recent LLM, Claude 3 Opus, outperforms earlier versions of the Claude models on all bias factors that they measures, including age, nationality, religion, gender and race: Anthropic, ‘The Claude 3 Model Family: Opus, Sonnet, Haiku’ (20 June 2024).

aiming for a high grade may be better off applying that labor to learning the subject material directly. On the other hand, the production of a mediocre, yet passable, essay appears well within reach of the current technology. This alone should elicit reflection among legal educators. Its impact may require a shift in assessment methods, increased digital literacy education for students and educators, and the introduction of guidelines on AI-usage in an educational context.[80]

These lessons also extend into the court room and legal practice. Recent cautionary tales highlight some of the risks of incorporating this technology into a litigious practice. In New York, a lawyer was chastised by the court for producing a brief containing non-existent citations, all generated by ChatGPT.[81] In Vancouver, a lawyer was ordered to personally pay the costs of opposing counsel's time and effort researching the non-existent cases that ChatGPT had inserted into her client's notice of application. Whilst the court did not find that the lawyer had an intention to deceive the court, the judgment notes that

> [c]iting fake cases in court filings and other materials handed up to the court is an abuse of process and is tantamount to making a false statement to the court. Unchecked, it can lead to a miscarriage of justice.[82]

Judiciaries around the world have also begun to recognise and respond to both the promise and perils of the technology. John Roberts, the Chief Justice of the US Supreme Court, has predicted judicial work, particularly at the trial level, would be significantly affected by AI, and noted its proponent's claims that it could increase access to justice.[83] Many judiciaries have released formal guidelines for both their judicial staff and lawyers appearing before the court.[84] Whilst the guidelines suggest certain appropriate uses, including summarisation and administrative tasks, they typically raise a set of common concerns, many of which have been highlighted in this study. These include the risk of hallucinations in generated content, along with a reminder to practitioners that the use of LLMs does not relieve them of their professional obligations not to mislead the court or other parties. The lack of privacy and concerns over how inputted data is used are also a major concern, and necessarily limits the scope of tasks that can be performed on privileged or sensitive case work. Practitioners are also warned about the models' bias, as discussed above. Indeed, the UK

[80] This is already the subject of much research. See, eg, Chung Kwan Lo, 'What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature' (2023) 13(4) *Education Sciences* 410.

[81] Weiser and Schweber (n 4).

[82] *Zhang v Chen* [2024] BCSC 285.

[83] John Roberts, '2023 Year-End Report on the Federal Judiciary'.

[84] Including in New Zealand (Courts of New Zealand, 'Guidlines for Use of Generative Artifical Intelligence in Courts and Tribunals' (7 December 2023)), Australia (Supreme Court of Victoria, 'Guidelines for Litigants: Responsible Use of Artificial Intelligence in Litigation' (6 May 2024); Australasian Institute of Judicial Administration, 'AI Decision-Making and the Courts. A Guide for Judges, Tribunal Members and Court Administrators' (June 2022)) and the United Kingdom (Courts and Tribunals Judiciary, 'Artificial Intelligence (AI) Guidance for Judicial Office Holders' (12 December 2023)).

guidelines explicitly note that the current LLMs appear to hold a 'view' of the law based heavily on 'US law'.[85]

VIII Limitations

This study had several important limitations. The results of this study are the product of a narrow set of methods and prompt styles. A small sample size was used. The clear absence of footnotes in the LLM-produced essays may have had an oversized influence on the way the essays were graded. The nature of conversational prompts and the non-determinism of the GPT-4 model means that there are many different ways by which the models can be prompted and by which the output can be combined.[86] Other systems, such as retrieval-augmented generation, are being developed to address the hallucination and citation shortcomings of the models.[87] In 2023, new models, systems and research into their use, which are all likely to improve output performance, are being rolled out at pace. A Google Scholar search reveals tens of thousands of results on the term ‘GPT-4’ in the less than 12 months since its release. Given the attention, the advances and breadth of possible deployment techniques, it is possible that a system of other long-form methods, prompting strategies and LLM models could be developed to produce content of significantly higher quality in the near future.

Aside from the generation of new, original content, the models can also be used in many ways to supplement existing writing, such as generating ideas for students, improving grammar, or editing first drafts. These less interventionist methods should also be on the radar of legal educators, but whether their use by students in that way is a concern, or potential boon, is another question altogether.

E *Future Work*

As noted above, this work is limited in scope and sample size. The big question remains, namely: What is the true capability of the models to produce persuasive legal writing, and longer form and persuasive content more generally?

To properly address this questions, future work may include:

1. A larger, systematic study, accounting for a broader range of prompt styles, essay questions, variance between graders and the non-determinism of the model output;

[85] Courts and Tribunals Judiciary (n 84) 3.

[86] See our explanation of 'non-determinism' (n 49) and 'combinatorial explosion' (n 64).

[87] Retrieval Augmented Generation (RAG) employs the LLM as a natural language interface to the bespoke information contained in an external database. Jaromir Savelka et al (n 12), found that augmenting a prompt to GPT-4 with relevant sentences from a case law database improved the quality of explanation of terms in legislation. A similar system could append, for example, sections of text from a curriculum reading list to an LLM prompt with a view to improving citation accuracy within an essay.

2. An analysis of other assessment task types (such as fact-based advice questions) and subject matter (such as substantive law); and/or
3. The use of other production methods that may improve model output, such as retrieval augmented generation, or other LLMs besides GPT-4.

Acknowledgements

We thank John Sebastian, Julian Webb, Shaun McVeigh and Dale Smith for grading the essays and providing comments. Thanks also to Jeannie Paterson for her helpful feedback on the manuscript.