Persuasive Legal Writing using Large Language Models

Abstract

Persuasive writing is an essential skill for a lawyer. Budding lawyers hone their persuasive writing skills during their studies, in part through essay-style examinations. However, Large Language Models (**LLMs**) have recently proved adept at a broad range of language tasks, including many of the elements essential to persuasive legal writing. They could undermine the utility of many existing forms of law-school assessment by allowing students to generate essays artificially.

We evaluate whether OpenAI's GPT-4 can produce essay-style answers to a post-graduate law school exam on legal theory. In doing so, we also develop and compare two methods for producing long-form content of more than 750 words from LLMs, which we call 'Accordion' and 'Recursive'. We compare the performance of the GPT-4 output against essays written by actual high-performing honours students by having them blind-graded by human graders using the subject's examination rubric. We measure the differences in performance between the artificially generated essays and the students, and between the essays produced by the two methods.

We find that GPT-4 can produce long form persuasive legal writing which is of a passable grade, but that there are significant challenges producing higher quality content.

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Motivation

Lawyers write to persuade. Their writing can take many forms, from a written submission made during a court case, a judgment, inter-partes correspondence between solicitors, 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. If that occurs, law students may not learn persuasive writing skills and their grades may not reflect their true abilities in this critical area as they step into the profession.

Powerful LLMs are already widely available for public, and often free, use. ChatGPT and its plug-ins, along with other similar architectures, have been deployed at pace in 2023 and are likely to proliferate. It would be a fair guess to presume they are already used by law students.

On the other hand, the capabilities of LLMs present a potential boon for the legal industry. Whilst they may not be ready immediately for unedited use in the courtroom [2], 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 such as inter-partes correspondence. This automation is likely to decrease operating costs within a firm, resulting in increased access to legal services, increased profit margins, or both.

It is therefore imperative that benchmarks are established which measure the persuasive legal writing ability of these LLMs. This will permit industry to gauge their utility, their potential use in practice and their liability risks, and for educational institutions to adapt their assessment methodologies.

1.2 Our Research

Our research asks the following question: **Can an LLM produce persuasive legal writing which mimics that of a high performing graduate law student?**

Specifically, we evaluate whether OpenAI's GPT-4 can produce essay-style answers to a post-graduate law school exam on legal theory and, if so, to what extent. In doing so we also:

i. Compare the performance of the GPT-4 output against actual high-performing student essays;
ii. Assess the performance of GPT-4 output in various sub-categories; and
iii. Develop and compare two methods for producing long-form content from LLMs, which we call 'Accordion' and 'Recursive'.

We produce four essays across two essay topics from the exam of Legal Theory, a graduate law class at the University of Melbourne [3]. For each topic we produce two essays - one with each method. We have these essays blind graded by experienced graders against the essays written by honours students. In addition to a final mark, the essays are graded against a rubric comprising three sub-categories. The sub-categories measure organisation and reasoning, factual knowledge, and the quality of the analysis and reflection. We measure the differences in performance between the artificially generated essays and the honours students, and between the essays produced by the two methods.

The GPT-4 output, whilst still on average passing the exam, performed worse than the students. As expected, the output was generally well structured and argued, exhibiting better performance in that criterion than in the other criteria. Given the well-known tendency of LLMs to hallucinate, it was unsurprising that it did not exhibit strong factual understanding or accurate knowledge. But despite suggestions in some work that LLMs could exhibit creativity, we did not see any elevated performance in the measure of critical analysis and originality, compared to the other criteria. We hypothesise that the muted performance in this category may be the result of the prompt engineering we conducted in order to curtail hallucinations. We did not observe any difference in performance between our long form production methods. The comments on the GPT-4 essays showed greater negative sentiment than the comments for the student essays.

1.3 Contribution

We make the following contributions to the literature:

1. We show that LLMs can produce long form persuasive legal writing content which is of a passable grade, but that there are significant challenges producing higher quality content;
2. We propose two methods for producing long form content from LLMs and begin a discussion about their possible advantages and use-cases;
3. We propose an assessment rubric which can be extended for use on other generated, long-form persuasive writing.

1.4 Outline of Thesis

We first provide an overview of the background literature in Chapter 2. We explain the rise of Large Language Models and why they may be suitable for the production of persuasive legal writing. We also identify the mixed results of recent experiments in which LLMs have been used to generate similar legal and essay-style writing. Finally, we explore the absence of literature on the using these tools to generate long-form content.

In Chapter 3 we explain the 'Accordion' and 'Recursive' methods for producing long-form content from LLMs we will be using in this study. In Chapter 4 we look at the evaluation methods we are using to evaluate them, including details of the human essays being used as a benchmark and the grading rubric. In Chapter 5 we provide our results. Finally, in Chapter 6 we place the results or our research into a broader context.

Chapter 2 Background

2.1 Large Language Models

Language models, at their simplest, predict the likelihood of a sequence of words [4]. By predicting the likelihood of a word sequence, a diverse range of downstream language tasks can be performed. There are many types of language models. Statistical language models have been in use for decades [5]. However, pretrained language models built upon the transformer architecture with self-attention mechanisms [6] have proliferated in recent years. They have several distinct advantages. These models are self-trained on large quantities of unlabeled corpora and fine-tuned for a specific task. This eliminates the need to develop expensive, labelled corpora, and allows much greater volumes of knowledge to be consumed during training. Their transformer architecture also allows for parallelization, which greatly reduced training time and cost. Their development, beginning in the late-2010's, resulted in notable performance gains across many natural language task benchmarks [7,8].

Research found that upward scaling pretrained language models resulted in predictable performance gains, including surprising 'emergent' abilities [9]. These scaled pretrained language models were coined 'large' language models. LLMs were several orders of magnitude larger than the early brethren and were trained on massive corpora. For example, whilst the original $\text{BERT}_{\text{LARGE}}$ model contained 340 million parameters [7], GPT-3, released only three years later, contained 175 billion parameters. (A 'parameter' of a language model is a variable, or weight, within the neural network architecture.) GPT-3 was trained on nearly a trillion words from the CommonCrawl dataset (a freely available repository of petabytes of webpages, collected from the web since 2008), books and the English-language Wikipedia. It was able to solve substantially more complex tasks than smaller models which had come before it [10].

In this study, we used OpenAI's latest publicly accessible LLM, GPT-4. Due to 'the competitive landscape and safety considerations', the architecture and training set of GPT-4 are not publicly available [11]. Nonetheless, some relevant information can be gleaned from its technical report and system card [11] and papers on earlier GPT iterations [8,10].

GPT-4 is trained in two stages [11]. It is first trained to predict the next word from a large dataset of text. It is in this stage that its underlying 'knowledge' is encoded into its parameters. Whilst the exact content of the training data is not publicly available, it is likely to include at least the massive corpus

used when training GPT-3, as noted above. Relevantly to this study, this is likely to include, for example, the Wikipedia entries on well-known legal theorists and their ideas. It is also likely to include many more related documents scraped from the web, such as student essays, blogs 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) [11]. In this stage, the model is further trained so that the knowledge it has learnt in its original training stage can be delivered in a format that is preferable to a human user and aligned with their intent. It is specifically optimized for dialogue with a human user [4]. This may include finetuning to follow its user instructions, whilst also minimizing the risk that it produces harmful content [11]. 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 [12].

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.

2.2 Persuasive Legal Writing

In this study, we want to apply GPT-4 to the task of generating persuasive legal writing. We define persuasive legal writing as *'text written in the legal domain for the purposes of persuasion.'* This writing style may also be known as written advocacy or legal argumentation. Persuasive legal writing can take many forms, from a written submission made during a court case, a judge's judgments, inter-partes correspondence between solicitors, 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.

This thesis is not a philosophical examination of the essence of persuasive legal writing. Nonetheless, it is necessary to consider the constituent elements of persuasive legal writing. That will allow us to conduct a fine-grained evaluation of the LLM's performance and make claims about generalizability beyond the university essay. To this end, we have conducted a brief survey of the literature regarding advocacy, argument and persuasive prose. This survey has revealed several common elements which contribute to effective persuasive writing in the legal domain. We discuss these themes below and identify how they map onto the marking rubric we will be using to evaluate the LLM output in this study. We also identify existing studies on the performance of LLMs in each of the constituent elements and identify the strengths and weaknesses of LLMs that we expect to encounter in this study.

2.2.1 Structure and Argument

Persuasive writing must be well structured. "Structure is important," notes Davies J [13] in a guide to persuasive written advocacy. "The document should provide an easy road map for the reader to follow so that the reader from the outset is able to follow the significance of what he or she is reading." She warns that "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 2.3, 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, as noted in Section 2.4, the ability of the models to produce longer form content has not been properly explored.

Persuasive writing must also present an argument. This is closely related to the ability of the LLM to reason. The Stanford Encyclopedia of Philosophy states that argumentation can be defined as "the communicative activity of producing and exchanging reasons in order to support claims or defend/challenge positions" [14]. Using this definition, arguments are comprised of 'reasons'. In an essay there may be one 'reason' per paragraph, presented as a topic sentence, an analysis and a concluding sentence.

Recent studies have shown GPT models are capable of performing various reasoning tasks [15,16]. 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 [17].

Several papers have explored the *legal* reasoning capability of LLMs. The results have been mixed. Explicit tests of GPT's ability to answer logic problems based on synthetic statutes was explored by Blair-Stanek et al [18]. They found that the model outperforms previous benchmarks, but still makes clear errors. As also noted in Section 2.3, several papers have explored the models' performance on law school exams. Good performance on these exams requires a student to analyse 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 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.

2.2.2 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. Credibility is akin to the writer's 'reputation', which a former Justice of the High Court of Australia, Kirby J [19], called an advocates 'most priceless possession'.

In a study by Savelka et al [20], an LLM was tasked with explaining how a key term from US statute was used in caselaw. The authors found that, despite responses appearing highly plausible, detailed analysis uncovered limitations in the factual accuracy of the explanations. Studies across other disciplines have also found that, when asked to cite sources, LLMs commonly either fabricate sources entirely, or conflate multiple sources into an original hybrid [21,22].

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.

2.2.3 Critical Analysis and Original Reflection

Our review suggests that 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 [23] suggests that "persuasion calls not only for mastery of the materials, but also for an element of constructive imagination and boldness of approach". In Law and Literature, Mr Justice Cardozo [24] 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 sincerity and fire, or the mnemonic power of alliteration and antithesis, or the terseness and tang of the proverb and the maxim."

LLMs have been shown to exhibit some form of empathy, creativity and reflective nous along these lines. Ayers et al [25] posed a series of questions from patients about medical issues on a social media forum and had ChatGPT produce answers to them. The chatbot answers were preferred over those of actual physicians and rated higher in empathy. Haase et al [26] implemented a test requiring participants to generate novel uses for a range of everyday objects. 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 [27] found that GPT-4 can produce reflective writing. These studies show that there is a reflective, creative and imaginative streak in GPT models.

2.3 Recent applications of LLMs to Essay Writing and Law Exams

A number of studies have examined the ability of an LLM to produce short essays. Yeadon et al [28] explored the capability of a GPT-4 predecessor model, text-davinci-003, to produce short essays for a first year university subject called 'Physics in Society'. Despite being an earlier model, the essays received first-class grades. Herbold et al [29] produced hundreds of high school essays using a range of OpenAI models and had them scored against non-native English speaking students. The GPT-4 essays received the highest grades, followed by those from GPT-3.5 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 Passes the Bar Exam' by Katz et al [30]. These findings were touted by OpenAI when it released GPT-4 [31]. 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. Subsequent papers have queried some of the methodology used in the paper [32]. The 90th percentile claim nonetheless suggests the model is very capable of passing difficult legal examinations and exhibits a substantial depth of 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, but nonetheless also suggest at least a passable ability of the models to perform legal reasoning. Choi et al [33] studied GPT-4's performance on a spread of law school exams from the University of Minnesota. The model showed only average (C+) performance, akin to "a bright student who never made it to class." Blair-Stanek et al [34] similarly assessed performance on University of Maryland law school exams, showing mixed results but uniformly below average. These results, again, suggest a decent ability to perform legal reasoning, but leave room for improvement.

2.4 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 (in contrast to much work on the *summarization* of long form content). Mirowski et al [35], acknowledging that language models lack long-range semantic coherence, built a system which generates coherent scripts and screenplays using prompt chaining. They used crowd-sourced feedback to evaluate the quality of the output. However, this was not a study of essay-style writing and further interest in the problem seems muted.

For the purposes of this paper, we consider 'long-form' content to be in excess of approximately 750 words. **Table 1** provides an indication of the length of essay-style content produced in a range of relevant papers published in the last 12 months. We have not identified papers which have explored the ability of LLMs to generate essay-style responses greater than about 600 words.

Table 1: Indicative word count of GPT produced content evaluated in recent, related studies

Paper	Text	Word Count
Liu et al 2023 [36]	Argumentative essay generated by GPT-4	300 – 500 words
Yeadon et al 2022 [28]	Short-form physics essay generated by ChatGPT	300 words
Herbold et al 2023 [29]	Argumentative student essays written by ChatGPT	200 words
Li et al, 2023 [27]	Student reflections written by ChatGPT	~ 300 words average
Katz et al, 2023 [30]	GPT-4 answers to MPT long-answer section on bar exam (as appended to paper)	~ 600 words

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. It is a 'hard' upper limit inherent in the design of LLMs. 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. **Table 2** shows context window limits of various LLMs.

Table 2: Context window size of recent LLMs

Model	Context Window Size (tokens)
GPT-2	1,024
GPT-3.5-turbo	4,097
GPT-4	8,192
GPT-3.5-turbo-16k	16,385
GPT-4-32k	32,768
Anthropic Claude	100,000
GPT-4-Turbo	128,000

The second factor is the model's ability to maintain optimum reasoning performance over a long context window. The table above shows a clear trend towards larger context windows. Some, such as Anthropic's 100,000 token Claude model, may seem adequately large to reason over even very long writing. However, Liu et al [37] 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. Therefore, the actual length over which an LLM can reason with *optimum* performance may be much smaller than the full context-window.

There are also practical limits on how long LLM output can be in a single inference. LLMs will stop generating when they produce a 'stop generating' token (or similar stop sequence). When, or if, this is produced will vary from model to model and depend upon its training. The RHLF training phase of GPT-4 will have fine-tuned the model to produce content of a certain average length. Preliminary testing by the authors 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.

In summary, each LLM will 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. This will be either a hard upper limit due to the model context window, or an effective upper limit due to degraded performance or a practical refusal to produce long content. It is therefore necessary to develop and evaluate different techniques to produce longer content by combining content produced in multiple inferences.

Chapter 3 Method: Long Form Production

3.1 Overview of Methods

We explore two methods to produce coherent, long form content from an LLM. The methods are 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 [38].

To that end, both methods are based on an 'outline' that has been independently generated by the LLM from the essay topic. The outline splits the essay into 'sections'. These can be anything from a short paragraph to several hundred words. Both methods produce the full text of one 'section' at a time. These sections are concatenated to form the final product.

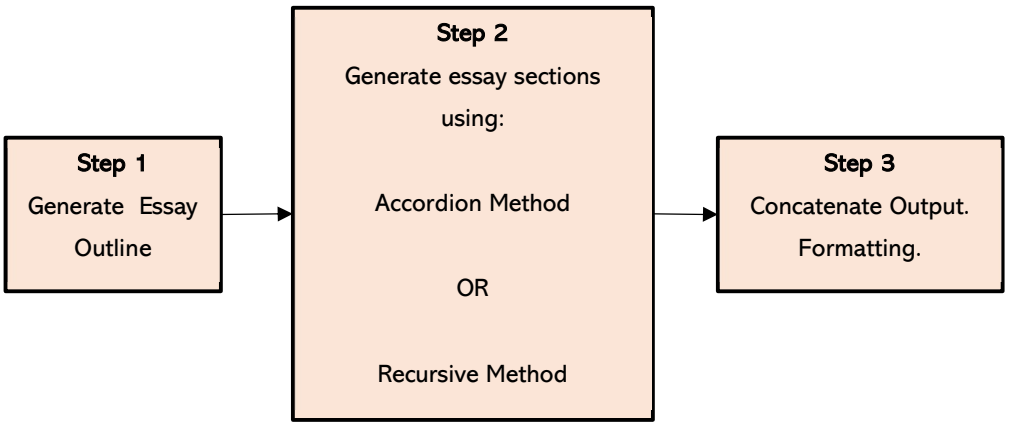

Figure 1: Overview of the process used to produce long form content from an LLM.

The difference between the methods is whether the previously produced text from earlier sections is included in the prompt for a subsequent section.

Method 1 (Accordion) does not include any earlier text produced by the model. It is akin to simply producing the content independently and concatenating the output. The model is 'blind' to all previous content produced by it. Because the prompt length remains static, this method could potentially have significantly cheaper inference cost and be used by smaller models with better reasoning performance.

Method 2 (Recursive) includes in every prompt all text previously produced by it. The model can 'see' all previous section text. However, the prompt length grows longer with each new section produced, which leads to high inference cost, the need for larger models and a potentially reduced reasoning performance (as shown by [37]). This method is akin to using GPT-4 in the ChatGPT web interface in which the entirety of the conversation and previously produced text is available to the model (subject to being within its context window).

The methods differ in the amount of earlier content they can 'see' when producing the next section of the text. We can hypothesise that this impacts their ability to produce coherent and cohesive output. We can make this hypothesis because 'seeing' the text from earlier sections may permit the production of subsequent sections which:

i. Include cohesive devices across sections, such as:
    a. Transition sentences and words which 'glance backwards' to the previous section;
    b. The use of terms which are defined in earlier sections;
    c. Anaphora in which the antecedent occurs in the earlier section.
ii. Have the same writing style or rhythm as earlier sections;
iii. Do not repeat phrasing, ideas or evidence from earlier sections;
iv. Are not inconsistent or contradictory with earlier sections.

However, including additional content in the prompt (as in Method 2), increases the cost of the inference, requires the use of models with larger context windows and may reduce reasoning performance [37].

The extent to which any of these differences are actually exhibited, and in which direction, is unknown. This thesis provides some initial experimental results aimed at addressing these questions.

3.2 Prompt Engineering

The methods described in the following sections use prompts which were developed by us. We iterated through many prompts before settling on those described below. We experimented with many prompt combinations and evaluated the quality of the LLM output of each combination, and then iterated. Evaluation of the interim output was conducted with reference to the Legal Theory evaluation criteria identified in the next section. It was conducted manually by the principal author, who is a qualified solicitor, but has not recently undertaken the Legal Theory course and has never taught in a law school.

Trends in the effectiveness of the different prompt styles were noted but no systematic engineering was conducted. Google searches were used to spot check the factual accuracy of some citations and sourcing produced by the LLM. This was akin to the process which a bright but non-committed student might undertake when using a tool such as ChatGPT, in which they iterate through various prompt combinations to find the best output in the limited time available to them during an exam window.

Our takeaways from the prompt engineering phase are that producing content which refers to and is dense with targeted citations to specific content from the course material is difficult. Trying to steer the model to do so increases the chance it will over-correct and hallucinate. Validating the accuracy of citations is significantly time consuming. Our detailed observations from that prompt engineering process are discussed in **Appendix 1**.

3.3 Essay Outline

Before either method is used, an essay 'outline' is independently produced, based on the essay topic and other information from the persona. The outline is produced by GPT-4 using its 'functions' ability, which returns output in a structured JSON format [39].

The outline is generated using a prompt template which contains the following components:

i. **background**: the background of the task;
ii. **persona**: the persona to be adopted by the LLM;
iii. **task**: specify exactly the output required, including options for the number of sections to produce, and whether the essay should agree or disagree with the topic;
iv. **references**: an optional input designed to steer the substantive content of the essay, such as what texts, authors or ideas should be referenced;
v. **topic**: the actual essay / exam question.

An example of the actual prompt used to produce the final essay is shown **Figure 2** for the 'judicial power' essay topic. Each component of the prompt is marked by square brackets and the subscript label.

```
[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]
```

Figure 2: Prompt used to generate an outline for the essay topic on 'judicial power'.[1]

An example outline output from the outline function is shown in **Figure 3** below.

```
{
    "sections": [
        {
            "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."
        },
…
{
```

Figure 3: An example outline generated by the outline prompt for the 'judicial power' essay.[2]

3.4 Common Prompt

Both methods generate the first section of the essay using an identical prompt. This prompt is a template comprised of the following components:

i. **background**: the background of the task (identical to the outline component of the same name);

[1] The components of the prompt template are marking in square brackets with a subscript label.
[2] The output format is JSON, which permits easier post-processing when automating essay generation.

ii. **persona**: the persona to be adopted by the LLM (identical to the outline component of the same name);
iii. **task**: specify exactly what is required in the output;
iv. **outline**: the entire outline from the outline function, described in the above section, converted to a string;
v. **section_length**: the length required of the present section;
vi. **additional_instruction**: any further tweaks required to the instructions during prompt engineering;
vii. **trigger**: the final part of the prompt, specifying which section to produce.

An example of the common prompt used to generate the introduction section for the essay on 'judicial power' is shown **Figure 4**. Each component of the prompt is marked in square brackets and the subscript label. The full text of the outline has been truncated in this figure.

```
[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]
```

Figure 4: The common prompt used by both methods to generate the first section of the essay on 'judicial power'.[3]

3.5 Method 1 (Accordion)

In Method 1 (Accordion) the LLM is provided with the initial prompt (above) and instructed to write the first section. The output is stored. The process is repeated *n* times until all sections are written (where *n* is the number of sections generated by the outline), each time changing only the title of the section to

[3] The components of the prompt template are marked in square brackets with a subscript label. The full text of the 'outline' has been truncated.

be produced in the 'trigger' component of the prompt template. Each section is produced independently, based only on the summary outline of the essay. The output is concatenated to produce the final essay. In other words, the model is 'blind' to the previously produced content.

This process is illustrated in **Figure 5**.

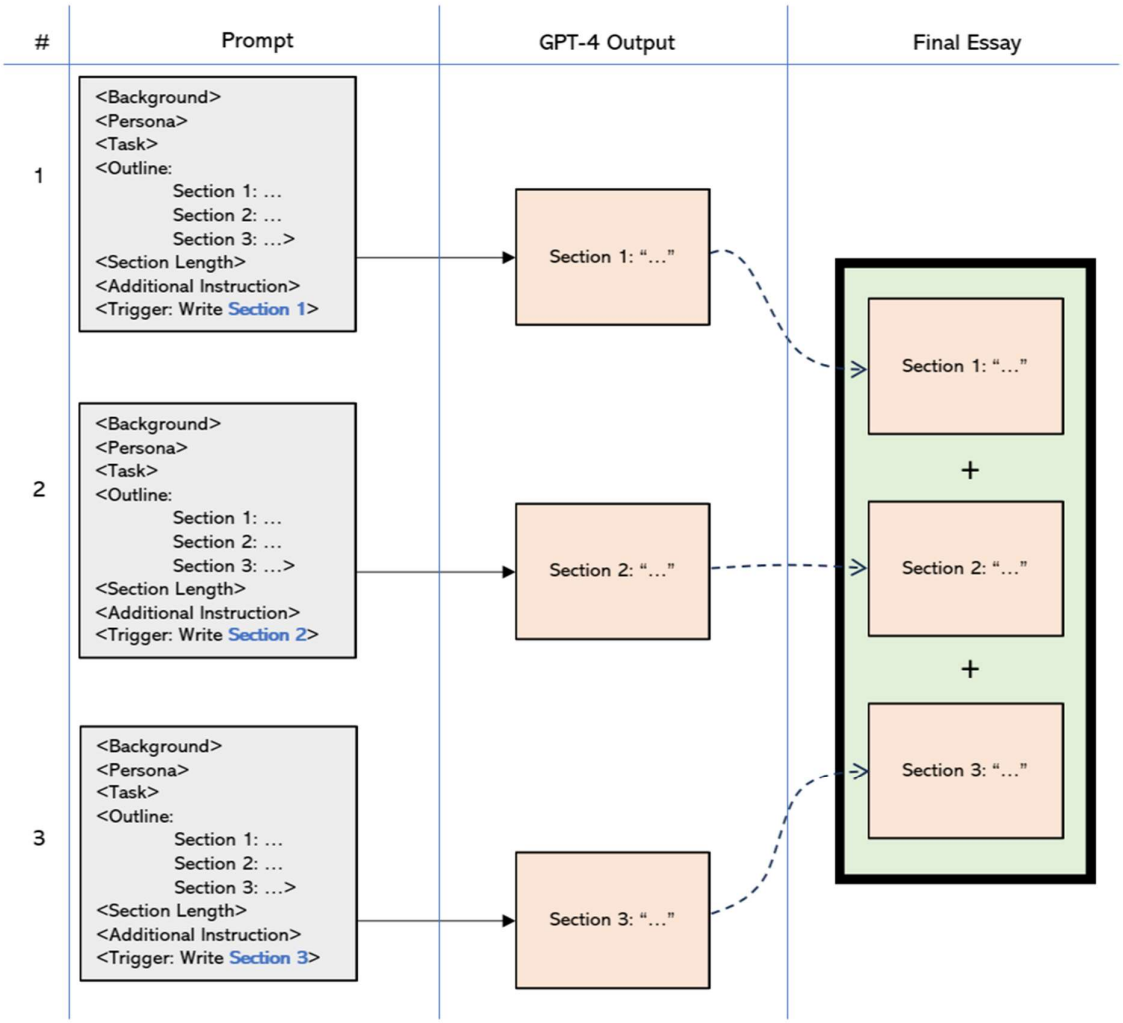

Figure 5: The Method 1 (Accordion) process, illustrated.

We refer to this method as 'Accordion' because it mimics the html accordion element. This element has vertically stacked headings (akin to our outline summaries) which expand to reveal detailed content when clicked (akin to our produced section content), and then collapse (or disappear) when another element is expanded. The html accordion element is illustrated in **Figure 6**.

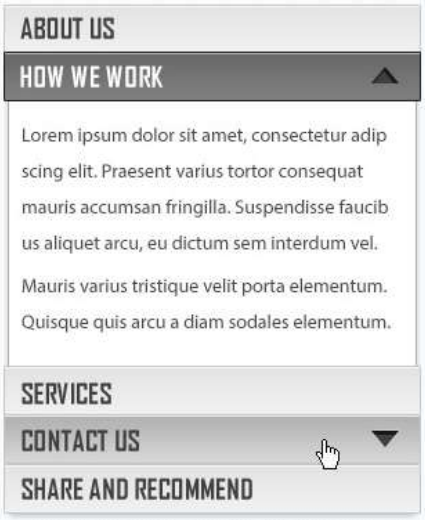

Figure 6: Example of a html accordion element [40].[4]

3.6 Method 2 (Recursive)

Similarly to Method 1, in Method 2 (Recursive) the LLM is provided with the initial prompt (above) and instructed to write the first section. That output is stored. On the next iteration, the content produced from the first production is appended to the end of the prompt, as well as a further 'nudge' asking the model to produce the next section. The nudge prompt is shown in **Figure 7**.

```
[In accordance with the earlier instructions, continue the essay by writing the next section only:
'The Concept of Law and Judicial Power' nudge]
```

Figure 7: The 'nudge' prompt used to trigger the LLM to produce the next section of the essay, used in Method 2 (Recursive)

This continues through *n* sections of the essay, with the prompt growing in length as more content is produced and appended. The LLM can 'see' all the previous content produced by it. (This assumes that the total length of the input prompt, including all previously produced text, is within the size of the model's context window. In this case, using GPT-4 on an essay of less than 1,500 words, that is the case. If the essay length exceeded the context window, this method would not be suitable, or would need to be amended to include, say, only the outline and the most recently produced sections.)

This process is illustrated in **Figure 8**, below.

[4] Method 1 is named after the accordion element because it expands the section summaries into more text, akin to the drop-down element.

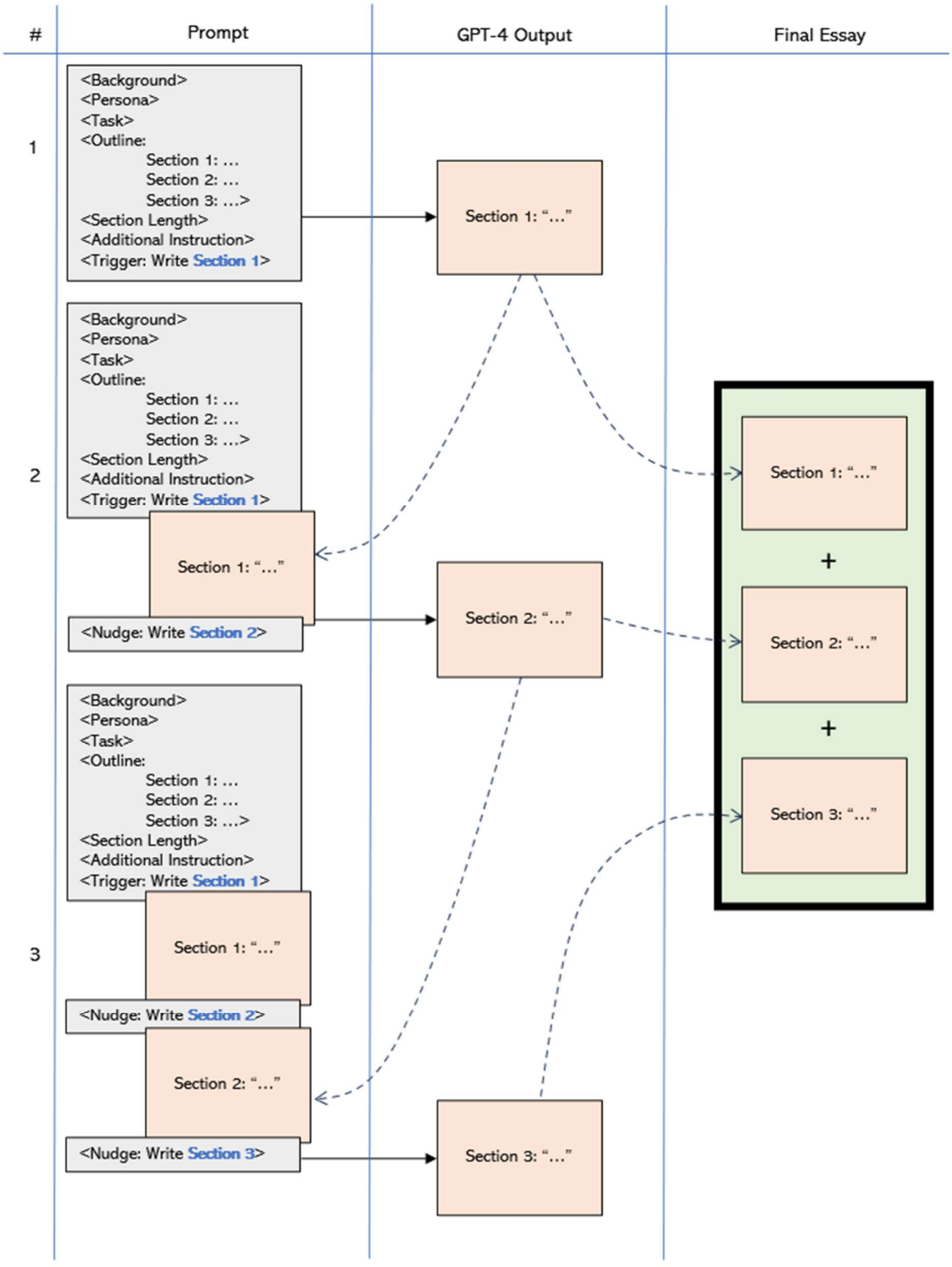

Figure 8: The Method 2 (Recursive) process, illustrated.

Chapter 4 Method: Essay Production and Evaluation

4.1 GPT-4

The LLM model to be used in this study is OpenAI's GPT-4 [31]. GPT-4 is available to the public through its chat interface and to developers through its API. It is used in a number of recent papers assessing an LLM's ability to produce law exam content, as well as studies assessing related abilities (such as [30,33,34]). We access the model through the OpenAI API. Automated techniques are coded in Python. The code used to produce the essays has been published and can be freely inspected.[5]

4.2 The Legal Theory Essay

The central artefact used in this study are essay questions from a Legal Theory examination. We evaluate GPT-4's ability to produce essays from that exam. Legal Theory is a postgraduate-level unit within the University of Melbourne's JD Program [3]. The subject explores various ways to think about legal concepts, institutions, processes, roles and values in the law. The unit spans limits of the law, obligations and rights of law, legal positivism and legal argument, amongst other matters.

The Legal Theory exam has several salient requirements. Students have three hours to write two essays from the four essay questions provided. The essay questions are open ended. They invite students to argue a position. They are typically one or two sentences in length. The total word count is 2,500 words across both essays. Whilst there is no minimum word count, high achieving students typically fill the quota, producing two essays each in the range of 1,000 – 1,500 words. Any citations must be clear and consistent.

4.3 Assessment Rubric

We had all the essays in our study (both human and GPT-4 generated) evaluated using the same rubric which a student's Legal Theory exam would ordinarily be graded against.

[5] https://github.com/dc435/LegalEssay_GPT

A grade of 'Excellent', 'Very Good', 'Good', 'Satisfactory', or 'Needs Improvement' were assigned to each of the following criteria:

i. **Argument and structure**: "Ability to develop an organised and reasoned response to the selected topic, justified by reference to relevant theorists and theoretical approaches." (**Argument and Structure)**
ii. **Knowledge and understanding**: "Knowledge and understanding of theoretical texts and arguments studied in the course relevant to the selected question." (**Knowledge and Understanding**)
iii. **Critical analysis and original reflection**: "Ability to critically analyse, compare, evaluate, situate and comment on theoretical arguments and accounts of law." (**Critical Analysis and Reflection**)

A final, overall grade is also given to the essay, out of H1 (excellent), H2A (very good), H2B (good), H3 (competent) or Fail.

These essays are a useful tool to answer our research questions. Each component of persuasive legal writing is addressed separately in the rubric, which will allow us to parse out the LLMs capabilities in each of the marking criteria separately.

4.4 LLM Essay Production

We produced four Legal Theory essays. Two exam questions were used. One essay was produced for each question using each method.

The exam questions were:

i. *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.*
ii. *Since the law is not morally-neutral, legal education must necessarily include an engagement with morality. Do you agree? Explain and justify your response.*

4.5 Student Essay Benchmark

The LLM essays were benchmarked against essays written by actual students who have previously undertaken the Legal Theory course.

Four essays were selected from two second-class honors students (a H2A and a H2B student) who had written on the selected topics. Second class honors is considered 'good' to 'very good' performance at a graduate level. It is not the top grade, 'excellent', but above the lower performing 'satisfactory' level. 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 student essays contained dense, pinpointed footnoting. Our testing during the development of the long-form method prompts confirmed that the LLM had difficulty producing content which could be reliably cited. This is a well-known shortcoming of LLMs [21,22]. This was the case whether the citation was 'in text' or contained in a footnote. There was therefore an obvious visual discrepancy between the students' essays and the artificially generated essays, in that the former contained footnotes on each page, and the latter did not. We were concerned that this visual discrepancy would signal to the graders that the LLM essays stood apart. We debated whether to edit the footnotes, move the citations in text, or delete them altogether. Such a course could have been justified, as this was not designed to be a study of the inability of LLM's to produce reliable citations (that being a well-publicized problem the subject of much research in the industry, and beyond the scope of this paper). However, we elected to maintain the footnotes in the student papers, because the graders needed to know if the writer was proposing an original idea, and whether the writer was relying on valid evidence. The deletion or amendment of footnotes could affect the substantive quality of the essay and its grading. We therefore retained the essay footnotes.

4.6 Grading

Four graders participated in the study. The graders were staff in the University of Melbourne Law School who had prior experience grading Legal Theory exams during the course using the same grading rubric. The graders were advised they were participating in a project evaluating a computational text analysis tool which required a comparison with human evaluations. (A copy of the instructions to the graders is included in **Appendix 2**).

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. We could not have four of the same topics given to the same grader, because the GPT-4 papers on the same topic are too similar, which may have aroused suspicion in the graders. **Table 3** shows the essay split amongst graders.

Table 3: The division of essays amongst the four graders.

Grader	Author	Topic
1	GPT-4 accordion	judicial power
1	H2A student	judicial power
1	GPT-4 accordion	legal education
1	H2B student	legal education
2	GPT-4 recursive	judicial power
2	H2B student	judicial power
2	GPT-4 recursive	legal education
2	H2A student	legal education
3	GPT-4 accordion	judicial power
3	H2B student	judicial power
3	GPT-4 accordion	legal education
3	H2A student	legal education
4	GPT-4 recursive	judicial power
4	H2A student	judicial power
4	GPT-4 recursive	legal education
4	H2B student	legal education

The graders were instructed to:

i. read each essay;
ii. rate the essay either Excellent, Very Good, Good, Satisfactory or Needs Improvement, in each category of the marking rubric (being the criteria from Section 4.3, namely Argument and Structure, Knowledge and Understanding, and Critical Analysis and Reflection); and
iii. Provide an overall grade for the essay of either H1, H2A, H2B, H3 or Fail.

They could also, optionally, provide additional comments on the essay. The detailed instructions are included in **Appendix 2**.

4.7 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.

To assess any differences in performance between GPT-4 and the students, and between the production methods, we looked at the median overall grades. Both the grade and individual criteria ratings are ordinal, so non-parametric methods are needed to test the significance of this difference [41]. We use the Mann-Whitney U test. It is a non-parametric alternative to the t-test. It tests the null hypothesis that the distribution of the samples in both populations is the same [42,43]. We used the Mann-Whitney U test to assess significance of differences in the median grades of GPT-4 essays compared to the student essays, and between the median grades of the Accordion essays and the Recursive essays.

To analyse any strengths of GPT-4 in certain sub-criteria, we used the Wilcoxon signed-rank test to compare the individual grade criteria against the median individual grade criteria for the GPT-4 essays. The Wilcoxon signed-rank test is the paired equivalent of the Mann-Whitney U test. It measures the median differences between two groups [44,45]. It is applicable when the sample size is very small [46], as it is here (N=8 GPT-4 essays). The null hypothesis for a one-sided test is that the differences in the paired values are symmetric around a number less than or equal to zero [47].

We perform a qualitative analysis on the comments by manual review and comparison. We also perform a sentiment analysis on the comment text. 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 [48]. We used VADER sentiment analysis tool [49]. VADER is a simple, yet popular, rule-based model for sentiment analysis. It has been used in previous studies including tools which auto-evaluate essay writing [50]. VADER produces a 'compound' score for any input text. This is a single measure of the sentiment of the text, normalized from -1.0 (very negative) to +1.0 (very positive). Our comments vary in length and number of sentences. To get a single sentiment value for each essay, we sententize the comment using spaCy's sententizer [51]. We then obtain the mean of the VADER compound scores for each sentence in the comment and compare those values for GPT-4 essays against the values for the student essays.

Chapter 5 Results

5.1 Criteria and Overall Grades

We received 16 graded essays from four graders. The results are shown in **Table 4**.

Table 4 - Individual Ratings and Overall Grade.[6]

				Individual Criteria Rating			
Grading packet	**Author**	**Type**	**Essay ID**	**Knowledge and Understanding**	**Critical Analysis and Reflection**	**Argument and Structure**	**Overall Grade**
1	gpt-4	accordion	essay10	Satisfactory	Satisfactory	Good	H3
3	gpt-4	accordion	essay10	Needs improvement	Needs improvement	Needs improvement	H3*
1	gpt-4	accordion	essay15	Needs improvement	Needs improvement	Satisfactory	Fail
3	gpt-4	accordion	essay15	Needs improvement+	Needs improvement+	Needs improvement	Fail
2	gpt-4	recursive	essay04	Needs improvement	Needs improvement	Needs improvement	Fail
4	gpt-4	recursive	essay04	Very good	Very good	Excellent	H2A
2	gpt-4	recursive	essay22	Satisfactory	Needs improvement	Needs improvement	H3
4	gpt-4	recursive	essay22	Satisfactory	Very good	Very good	H2B
2	student	H2A	essay07	Excellent	Excellent	Very good	H1
3	student	H2A	essay07	Satisfactory	Satisfactory	Good	H3
1	student	H2A	essay18	Excellent	Very good	Very good	H2A
4	student	H2A	essay18	Good	Good	Very good	H2B
1	student	H2B	essay12	Good	Very good	Good	H2B
4	student	H2B	essay12	Good	Satisfactory	Satisfactory	H3#
2	student	H2B	essay30	Good	Good	Very good	H2B
3	student	H2B	essay30	Good	Satisfactory	Satisfactory	H2B

The median Grade for the GPT-4 essays was H3. They achieved consistent grading of the same essay between graders (within one grade rank), except for 'essay04', which received a Fail from one grader and a H2A from another. The essay grades for the student essays matched or were within one grade rank of their original classifications, save one essay which was graded two ranks lower than its original classification. Some of this variance from the original classifications may be explained by earlier moderation of the original grades, which may have occurred naturally to adjust for the quality of the

[6] Some categories were manually adjusted to nearest fixed grading class, as follows: * 'Pass/H3' converted to H3; + 'Fail/NI' converted to Needs Improvement; # 'Pass' converted to H3

competing student essays when first assessed, or to adhere to a mark distribution policy. The median Grade received for the student essays in this study was H2B.

5.2 Grade Analysis

The minimum, median and maximum values for the GPT-4 and Student essays are shown in **Figure 9**.

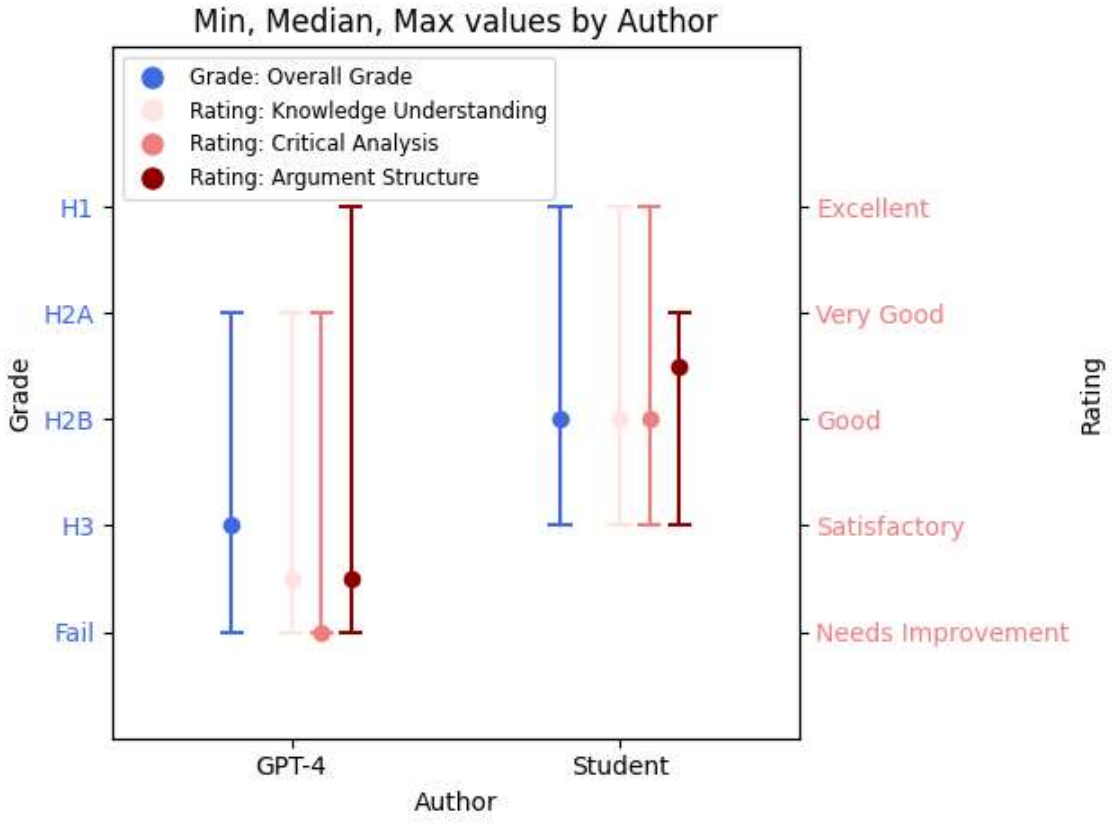

Figure 9: Minimum and maximum (denoted with '-') and median (denoted by 'o') values by author.

The student essays median grade (H2B) was one clear grade category higher than the GPT-4 essays (H3). The Mann-Whitney U test returned a p-value of 0.05, sitting right on the typical 0.05 threshold. The null hypothesis can therefore (just) be rejected. The difference between the GPT-4 and Student Overall Grade distributions is significant.

We conduct a Mann-Whitney U test to compare the Grades of the Accordion GPT-4 essays to the Recursive GPT-4 essays. We did not find any significant difference between the two methods.

Figure 9 also shows slightly elevated ratings for Argument and Structure for the GPT-4 essays compared to the other criteria. This suggests that the GPT-4 essays may perform slightly better at Argument and Structure than in the other categories. We compared the median of all three criteria against just the Argument and Structure criteria. The one-sided Wilcoxon signed-rank test returns a p-value of 0.041, slightly below the 0.05 threshold. The positive difference between the Argument and Structure criteria and the median of all the criteria is significant.

5.3 Comment analysis

We received optional comments from the graders for 14 of the 16 essays, 6 of which were comments for the GPT-4 essays and 8 of which were for student essays.

The 8 comments for the student essays were mixed, but generally constructive. The H2A 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 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 6 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'. They highlight the absence of references. The limited positive feedback was that one of the essays was 'clearly written', and another was 'well argued'.

We perform a sentiment analysis on the comments using VADER. The comment sentiment means are plotted in **Figure 10**. 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.

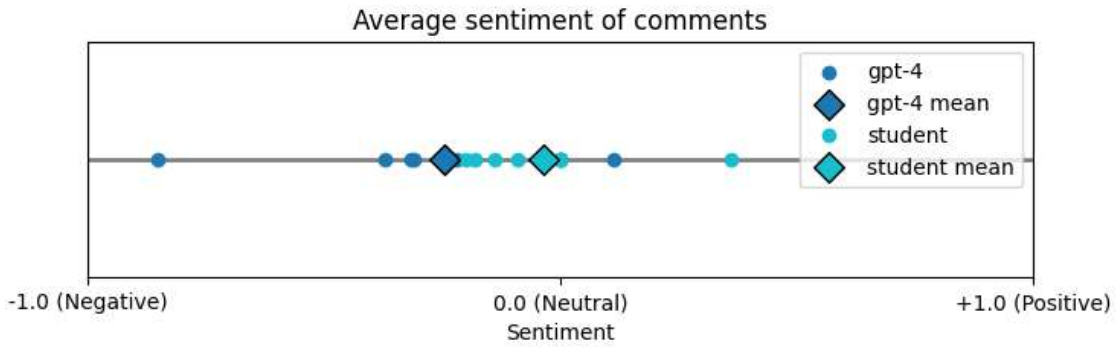

Figure 10: Average sentiment of comments

Chapter 6 Discussion and Future Work

6.1 Ability of LLMs to generate long form persuasive legal writing

Our results are consistent with existing literature from Choi [33] and Blair-Stanek [34] who in the academic legal context also found the models showed variable performance which was passable on average, but not excellent.

Our finding that the essays were slightly more proficient at Argument and Structure than at the other areas is to be expected, given the long list of 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 which studies in other fields suggested we might expect. This muted performance may have been the result of the decisions we made during the prompt engineering phase, which sought to curtail the output of the models so as to reduce the risk of producing content which was factually inaccurate or absurd. One the one hand this reduced the scope for the model to hallucinate, but may have muted the model's 'flair'.

6.2 Long-form production methods

We show that long-form production methods can be used to produce persuasive legal writing. Our paper begins a discussion on the possible methods to produce this type of content and identifies some potential trade-offs, such as inference cost and cohesiveness. However, we were unable to identify any difference in the output quality between the two methods. Because of the highly variable nature of the essay content, prompt style and non-determinism of the LLMs, a much larger sample size is likely to be needed to observe any statistically significant difference between these methods.

6.3 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. We showed that generating long form content using GPT-4 is difficult using simple prompts. A more complex, multi-step approach, such as the Accordion or Recursive methods we developed, is required to produce content at length. We showed that whilst concerns such as factual inaccuracies and hallucination can be quelled by tinkering with prompts, this also subdues the creativity and breadth of the content that is produced. Even after these tweaks, the essay cannot be guaranteed to be factually accurate, or even be guaranteed to pass. In summary, producing a decent output is challenging and requires a significant amount of manual tinkering. A student may be better off applying that labour to learning the subject material directly.

The stakes in professional legal practice are typically higher than for the production of a student essay. The threshold for factual accuracy and strong argument are much higher, because real livelihoods, freedoms and money are at stake. So whilst we have shown an LLM can produce a 'passable' student essay, a passable legal brief, submission or pleading is usually insufficient. The scope for its unedited use in real legal practice therefore seems limited (at least using the methods and model we adopted), outside of preparing first drafts, or other documents for internal use only.

6.4 Limitations

Nonetheless, legal educators and practitioners should continue to pay attention to this technology. Our results are the product of a very narrow set of methods and prompt styles. We used a relatively small sample size. 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. Other systems, such as retrieval-augmented generation, are being developed to address the hallucination and citation shortcomings of the models. In 2023, new models, systems and research into their use, which are all likely to improve output performance, are being rolled out at pace. In only 12 months since the release of ChatGPT, OpenAI's most advanced model went from the 4k token window gpt-3.5-turbo, to the just released 128k context window GPT-4-turbo model. A Google Scholar search reveals over 4,500 papers with a hit on the term "GPT-4" in 2023, an average of some 15 articles a day. 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 generation of new, original content, the models can also be used in many ways to supplement existing writing, such as generating ideas and editing existing content. 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 is another question.

6.5 Future Work

As noted above, our 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:

i. 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;
ii. An analysis of other assessment task types (such as fact-based advice questions) and subject matter (such as substantive law);
iii. The application of computational linguistic or automated metrics of cohesion and cohesiveness, in order to evaluate the difference in performance of the long form production methods; and/or
iv. The use of other production methods which may improve model output, such as retrieval augmented generation, or other LLMs besides GPT-4.

References

1 Michael R. Smith. *Advanced Legal Writing - Theories and Strategies in Persuasive Writing*. 3rd ed. 2014.

2 The New York Times. The ChatGPT Lawyer Explains Himself. 2023. https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html (accessed 9 August 2023)

3 The University of Melbourne. Legal Theory (LAWS50031) Handbook Entry. https://handbook.unimelb.edu.au/subjects/laws50031 (accessed 13 August 2023)

4 Zhao WX, Zhou K, Li J, *et al.* A Survey of Large Language Models. Published Online First: 2023. doi: 10.48550/ARXIV.2303.18223

5 Rosenfeld R. Two decades of statistical language modeling: where do we go from here? *Proc IEEE*. 2000;88:1270–8.

6 Vaswani A, Shazeer N, Parmar N, *et al.* Attention Is All You Need. Published Online First: 2017. doi: 10.48550/ARXIV.1706.03762

7 Devlin J, Chang M-W, Lee K, *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published Online First: 2018. doi: 10.48550/ARXIV.1810.04805

8 Radford A, Wu J, Child R, *et al.* Language Models are Unsupervised Multitask Learners. *OpenAI*. Published Online First: 2020.

9 Wei J, Tay Y, Bommasani R, *et al.* Emergent Abilities of Large Language Models. Published Online First: 2022. doi: 10.48550/ARXIV.2206.07682

10 Brown TB, Mann B, Ryder N, *et al.* Language Models are Few-Shot Learners. Published Online First: 2020. doi: 10.48550/ARXIV.2005.14165

11 OpenAI. GPT-4 Technical Report. Published Online First: 2023. doi: 10.48550/ARXIV.2303.08774

12 Zhou Y, Muresanu AI, Han Z, *et al.* Large Language Models Are Human-Level Prompt Engineers. Published Online First: 2022. doi: 10.48550/ARXIV.2211.01910

13 Justice Jennifer Davies. Effective and Persuasive Written Advocacy. 2013. https://www.fedcourt.gov.au/digital-law-library/judges-speeches/justice-davies/davies-j-20130807 (accessed 7 August 2023)

14 Argument and Argumentation. Stanford Encyclopedia of Philosophy. 2021. https://plato.stanford.edu/entries/argument/

15 Liu H, Ning R, Teng Z, *et al.* Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. Published Online First: 2023. doi: 10.48550/ARXIV.2304.03439

16 Espejel JL, Ettifouri EH, Alassan MSY, *et al.* GPT-3.5 vs GPT-4: Evaluating ChatGPT's Reasoning Performance in Zero-shot Learning. Published Online First: 2023. doi: 10.48550/ARXIV.2305.12477

17 Ellsworth P. *"Legal Reasoning" in The Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press 2005.

18 Blair-Stanek A, Holzenberger N, Van Durme B. Can GPT-3 Perform Statutory Reasoning? Published Online First: 2023. doi: 10.48550/ARXIV.2302.06100

19 Michael Kirby. Rules of Appellate Advocacy: An Australian Perspective. *ournal of Appellate Practice and Process*. 1999;227.

20 Savelka J, Ashley KD, Gray MA, *et al.* Explaining Legal Concepts with Augmented Large Language Models (GPT-4). Published Online First: 2023. doi: 10.48550/ARXIV.2306.09525

21 Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. Published Online First: 19 February 2023. doi: 10.7759/cureus.35179

22 Pride D, Cancellieri M, Knoth P. CORE-GPT: Combining Open Access research and large language models for credible, trustworthy question answering. Published Online First: 2023. doi: 10.48550/ARXIV.2307.04683

23 Honourable Mr Justice A. F. Mason. The role of counsel and appellate advocacy. *Australian Law Journal*. 1984;58.

24 Mr Justice Cardozo. Law and Literature. *Yale Review*. 1925.

25 Ayers JW, Poliak A, Dredze M, *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183:589.

26 Haase J, Hanel PHP. Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. Published Online First: 2023. doi: 10.48550/ARXIV.2303.12003

27 Li Y, Sha L, Yan L, *et al.* Can large language models write reflectively. *Computers and Education: Artificial Intelligence*. 2023;4:100140.

28 Yeadon W, Inyang O-O, Mizouri A, *et al.* The death of the short-form physics essay in the coming AI revolution. *Phys Educ*. 2023;58:035027.

29 Herbold S, Hautli-Janisz A, Heuer U, *et al.* AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. Published Online First: 2023. doi: 10.48550/ARXIV.2304.14276

30 Katz DM, Bommarito MJ, Gao S, *et al.* GPT-4 Passes the Bar Exam. *SSRN Journal*. Published Online First: 2023. doi: 10.2139/ssrn.4389233

31 OpenAI. GPT-4. 2023. https://openai.com/research/gpt-4 (accessed 13 August 2023)

32 Martínez E. Re-Evaluating GPT-4's Bar Exam Performance. *SSRN Journal*. Published Online First: 2023. doi: 10.2139/ssrn.4441311

33 Choi JH, Hickman KE, Monahan A, *et al.* ChatGPT Goes to Law School. *SSRN Journal*. Published Online First: 2023. doi: 10.2139/ssrn.4335905

34 Blair-Stanek A, Carstens A-M, Goldberg DS, *et al.* GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B. *SSRN Journal*. Published Online First: 2023. doi: 10.2139/ssrn.4443471

35 Mirowski P, Mathewson KW, Pittman J, *et al.* Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM 2023:1–34. https://doi.org/10.1145/3544548.3581225

36 Liu Y, Zhang Z, Zhang W, *et al.* ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. Published Online First: 2023. doi: 10.48550/ARXIV.2304.07666

37 Liu NF, Lin K, Hewitt J, *et al.* Lost in the Middle: How Language Models Use Long Contexts. Published Online First: 2023. doi: 10.48550/ARXIV.2307.03172

38 Cleaver M. 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 (accessed 1 October 2023)

39 OpenAI. GPT API Documentation - Functions Calling. https://platform.openai.com/docs/guides/gpt/function-calling (accessed 30 September 2023)

40 Wikipedia. Accordion (GUI). https://en.wikipedia.org/wiki/Accordion_(GUI) (accessed 30 September 2023)

41 Mircioiu C, Atkinson J. A Comparison of Parametric and Non-Parametric Methods Applied to a Likert Scale. *Pharmacy*. 2017;5:26.

42 Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Statist*. 1947;18:50–60.

43 Mann–Whitney Test. *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York 2008:327–9. https://doi.org/10.1007/978-0-387-32833-1_243

44 Schuff H, Vanderlyn L, Adel H, *et al.* How to do human evaluation: A brief introduction to user studies in NLP. *Nat Lang Eng*. 2023;29:1199–222.

45 Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1945;1:80.

46 Patwary MJA, Wang X-Z, Yan D. Impact of Fuzziness Measures on the Performance of Semi-supervised Learning. *Int J Fuzzy Syst*. 2019;21:1430–42.

47 Dror R, Baumer G, Shlomov S, *et al.* The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics 2018:1383–92. https://doi.org/10.18653/v1/P18-1128

48 Bonta V, Kumaresh N, Janardhan N. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *AJCST*. 2019;8:1–6.

49 Hutto C, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *ICWSM*. 2014;8:216–25.

50 Janda HK, Pawar A, Du S, *et al.* Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation. *IEEE Access*. 2019;7:108486–503.

51 SpaCy. Sentencizer. https://spacy.io/api/sentencizer

Appendices

Appendix 1 – Prompt Engineering Observations

Appendix 2 – Instructions for Essay Graders

Appendix 1 – Prompt Engineering Observations

In this Appendix we discuss our observations from iterating through the different prompts used to produce the summary outline and the essay sections of the long form techniques, and of reviewing the outputs from those prompts.

Background to prompt engineering

Rather than fine-tuning for a specific task, language tasks are solved on LLMs using prompts. A prompt is a text string which the language model probabilistically completes so that the output look like the strings it solved during its training phase.[1] Prompting can be thought of as a method to locate within a large language model a task that has already been learned by it.[2] Because prompts are the only way to guide these models to produce the 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.[3]

Prompt engineering is a rapidly developing field. Liu et al suggests the paradigm shift from 'pre-train and fine-tune' towards 'pre-train, prompt and predict' occurred as recently as 2021.[4] They provide a formal description of prompting. One method they discuss involves the development of a 'prompt template'. A prompt template has an input slot '[X]' and an output slot '[Z]'. Specifically in the case of GPT models (the model type to be explored in this study), a 'prefix prompt' is used in which the output slot appears on the end of the prompt template.

Many other prompting methods have been reported to increase the performance of LLMs. Li et al found that adding 'directional stimulus' to the prompt provided more fine grained guidance and control over the LLM.[5] Others have found that minor differences to the ordering of the prompts within the prompt window can affect performance.[6]

---

[1] Liu P, Yuan W, Fu J, *et al.* Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput Surv*. 2023;55:1–35

[2] Reynolds L, McDonell K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. Published Online First: 2021. doi: 10.48550/ARXIV.2102.07350

[3] Zhou Y, Muresanu AI, Han Z, *et al.* Large Language Models Are Human-Level Prompt Engineers. Published Online First: 2022. doi: 10.48550/ARXIV.2211.01910

[4] Liu P, Yuan W, Fu J, *et al.* Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput Surv*. 2023;55:1–35

[5] Li Z, Peng B, He P, *et al.* Guiding Large Language Models via Directional Stimulus Prompting. Published Online First: 2023. doi: 10.48550/ARXIV.2302.11520

[6] Liu Y, Zeng X, Meng F, *et al.* Instruction Position Matters in Sequence Generation with Large Language Models. Published Online First: 2023. doi: 10.48550/ARXIV.2308.12097

Our Prompt Engineering Phase

We began the prompt engineering phase by experimenting with prompts which mirror the simple instructions given to students undertaking the course. An example is shown in **Figure 1**. 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).

```
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.
```

Figure 1: Example of an early prompt which we experimented with whilst developing the prompt templates.

The output 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:

i. 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. This issue is discussed at length elsewhere in this thesis;
ii. They were often US centric. For instance, they would cite legal cases such as *Roe v Wade* and *Brown v Board of Education*;
iii. 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.

We then experimented with building up the prompt templates to further 'steer' the output. We added explicit instructions about the essay being written for a course in Australia and requiring a focus on theoretical arguments rather than empiricism.

But even after these corrections, the style of the output was still somewhat bland. The difference between the GPT-4 style and that of an honours student was marked. The honours student essays were much more targeted. They contained pinpointed references to academic papers and quotations and showed a command of the course material.

It was necessary to further 'steer' the GPT-4 model so that it would refer specifically to theorists, papers and quotations from the Legal Theory reading guide. An example of how we experimented with further steering of the output is shown in **Figure 2**. This prompt included references to the content from the subject reading guide for the chapter relevant to the essay topic.

```
… [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]  …
```

Figure 2: An example 'references' component of the prompt template which was experimented with. We 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 we used prompts which were very specific (such as the example in **Figure 2**), we 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. When it was steered in this manner, the output often hallucinated content from the sources that were included in the prompt. Whilst we did not undertake a systematic review of the possible hallucinations, 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 an author's point of view, confidently asserting the complete opposite position of that author.

The challenge, therefore, was to include some light steering which guided the output towards the desired themes, ideas and source material, but which did not overly do so. For the final essay productions, we included a prompt which mentioned a few of the key theorists which were focused on in the course, but did not specify any works or ideas of theirs. We found that doing so allowed 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.

We also identified that the ordering of the prompt components made a difference. We found that 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 the prompt was reiterated at the end of the prompt window. This is consistent with Liu et al.[7]

[7] Liu NF, Lin K, Hewitt J, *et al.* Lost in the Middle: How Language Models Use Long Contexts. Published Online First: 2023. doi: 10.48550/ARXIV.2307.03172

In summary, during the prompt engineering phase we learnt that:

i. GPT-4 is able to produce a decent, short essay from a very simple, short prompt;
ii. However, producing content which refers to and is dense with targeted references to specific content from the course material is significantly more difficult. Trying to steer the model in that direction increases the chance it will over-correct and hallucinate. Validating the veracity of that output is significantly time consuming. Any student wishing to use GPT-4 to produce persuasive legal writing in this manner might spend as long on this process as they might simply studying the material (or will, in any event, need to approach the task already having a good grasp of the course material).

Appendix 2 –Instructions for Essay Graders

Below are the instructions given to graders:

Background

We are conducting a study involving the analysis of responses to Legal Theory exam questions.

We require experienced legal theory teachers to evaluate exam answers.

The packets contain responses to the following essay questions:

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.*

You will be provided with a packet containing four essay responses. Each packet will contain two responses to Question 1 and two responses to Question 2. Each essay is given a unique number (XX) and the file is labelled in the form 'essayXX.docx'.

The exam answers have been edited so that their formatting is consistent.

Your Task

Please read each essay.

Rate the essay either Excellent, Very Good, Good, Satisfactory or Needs Improvement, in each of these categories:

i. Knowledge and understanding: Knowledge and understanding of theoretical texts and arguments studied in the course relevant to the selected question.
ii. Critical analysis and original reflection: Ability to critically analyse, compare, evaluate, situate and comment on theoretical arguments and accounts of law.
iii. Argument and structure: Ability to develop an organised and reasoned response to the selected topic, justified by reference to relevant theorists and theoretical approaches.

Provide an overall grade for the essay of either H1, H2A, H2B, H3 or Fail.

You may also provide additional comments on the essay. This is optional.

Please return the table below to us:

Essay #	**i. Knowledge and understanding**	**ii. Critical analysis and original reflection**	**iii. Argument and structure**	**Overall Grade**	**Comments (Optional)**
	(Options: Excellent; Very good Good; Satisfactory; Needs improvement)			*(Options: H1 - H2A - H2B - H3 – Fail)*	