Volume 6 (3) 2024 https://doi.org/10.5204/lthj.3637

# Using Generative AI to Identify Arguments in Judges' Reasons: Accuracy and Benefits for Students

**Paul Burgess, Iwan Williams, Lizhen Qu and Weiqing Wang**
Monash University, Australia

## Abstract

This study evaluates the effectiveness of generative artificial intelligence (GAI) in identifying and reconstructing legal arguments from judges' reasons in court cases, focusing on the practical implications for law students and legal educators. By examining the performance of two versions of popular Large Language Models – ChatGPT and Claude – across five recent High Court of Australia decisions, the study makes a preliminary assessment of the accuracy of LLM systems in replicating a skill essential for lawyers: identification of arguments and argument chains in judges' reasons. The methodology involves marking LLM-generated outputs with reference to both a sample answer and a detailed rubric.

Key findings reveal a significant variance in the accuracy of different LLMs, with Claude 3.5 markedly outperforming all others, achieving average grades up to 90 per cent. In contrast, ChatGPT versions demonstrated lower accuracy, with average marks not exceeding 50 per cent. These results highlight the critical importance of selecting the right GAI system for legal applications, as well as the necessity for users to critically engage with AI outputs rather than relying solely on automated tools.

The study concludes that while LLMs hold potential benefits for the legal profession, including increased efficiency and enhanced access to justice, for GAI use that may be carried out by a law student, the technology cannot yet replace the nuanced human skill of legal argument analysis.

***Keywords**:* generative AI; Large Language Models; arguments; education; judges' reasons.

## 1. Introduction

In this article, we demonstrate that Large Language Models (LLMs), as a representation of one form of generative artificial intelligence (GAI), have some capacity to identify and reconstruct the arguments used by judges in giving reasons to determine recent legal cases in the High Court of Australia. We also suggest that there can be benefits to junior lawyers (including students) from using artificial intelligence (AI) in this way. However, while these forms of technology are poised to revolutionise the legal industry and legal education, it is also clear that not all systems widely available at the time of writing perform at the same standard, and the nature of the benefit that follows from the various systems is commensurately limited. Although GAI technologies have been applied to build various legal applications and solve different legal tasks, to the best of our knowledge there is no study evaluating the capability of GAI on legal argument identification within the limitations related to the task in this article.[1] As a result, any potential benefit is linked to the choice of LLM system; an ability to critically engage with the LLM output thus remains critical. This means that, at least for the moment, the exclusive use of LLM systems as a replacement for the human skill of identifying arguments in judges' reasons is ill-advised.

[1] Kang, "Can ChatGPT Perform Reasoning?"; Wiratunga, "CBR-RAG"; Guha, "LegalBench."

We support these conclusions by testing the ability of two versions of both ChatGPT and Claude to accurately identify the chain of argument leading to the disposition of five recent matters in the High Court of Australia. ChatGPT and Claude have been chosen for this study because of their excellent performance in empirical studies regarding basic and advanced abilities in real-world applications.[2] By testing several versions of popular LLMs in a task similar to that in which law students may engage, we assessed the accuracy of the systems' outputs against a sample answer. The outputs were marked in accordance with a rubric that includes criteria such as the identification of the legally relevant conclusion and the arguments leading to the conclusion, together with locating the arguments within the given reasons. In other words, we treated the LLM systems' output like a student-created piece of work.

The study is limited in scope due to a familiar bottleneck in this field: the high cost of labour involved in the analysis of legal documents, which necessitates small numbers of annotators/assessors.[3] This represents a limitation on the conclusions reached. Nevertheless, while our study was limited, it serves as an initial benchmarking exercise, sufficient to draw some broad, but important conclusions about the role of GAI in legal education.

After marking the different LLM system outputs, the average grades (from two academics' blind marking of the outputs) ranged from 20 per cent to 90 per cent. It was clear from the distribution of grades that one LLM system – Claude 3.5 – outperformed all others. It was also clear that two versions of ChatGPT performed, on average, worse than Claude, with average marks of no more than 50 per cent. On this basis, we demonstrate that the choice of LLM system and version is crucial when it comes to replicating the core skills associated with being a lawyer. These findings provide vital guidance for practitioners, legal educators, and law students and give an indication of at least some of the changes and improvements that need to be made before the use of LLMs can completely revolutionise the legal industry.

We set out to answer the following questions: (1) can LLMs identify and extract legal arguments from court cases; and (2), to the extent that they can, how might this be useful in legal education? Importantly, we seek to answer these questions from the perspective of a 'typical' law student – one who cannot use complex prompting techniques and is instead limited to a one-shot prompt being submitted with a copy of the relevant judgment.[4]

With respect to the first question, we find evidence that some LLMs can reasonably accurately identify arguments given by judges in arriving at the conclusions that determine the issues in legal cases; however, some systems are not capable of doing so and the accuracy across systems varies considerably.

With respect to the second question, we argue that where LLMs *can* identify the arguments that are used by a judge in giving their reasons, there is benefit to students in using those LLMs in this way. However, while there is a significant gap between various systems' ability to identify arguments, the difference may not be immediately apparent to a layperson, as even the poorly identified arguments may superficially appear to be well-stated answers.

Thus, an ability to engage critically with those responses is necessary for students to establish whether the system output is accurate. This is especially crucial as our preliminary data suggest that,[5] by using a single prompt of the sort a 'typical' law student may use, the most well-known LLM system (ChatGPT) provides the least accurate responses (we tested two versions of the GPT models: GPT-4 and GPT4o) when contrasted with the less well-known Claude (3.0 and 3.5).[6] The combination of the potential skill-gap and the somewhat convincing outputs could result in significant detriment to the unskilled user.

In Section 2, we very briefly explain the importance of arguments (and their identification) in judges' reasons. We then provide some brief background detail about LLMs. Sections 3 and 4 relate to the steps we have taken to arrive at our results; we first set out the rationale for the approach that we have adopted before setting out the methodology. We then set out the results as well as responses to the two aspects of the research question.

[2] Zhao, "A Survey of Large Language Models."
[3] Savelka, "Can GPT-4 Support Analysis of Textual Data?"
[4] It has been established that more sophisticated prompting methods can elicit more accurate responses that one-shot single prompts. See, for example, Wang, "Towards Understanding Chain-of-Thought Prompting"; Wei, "Chain-of-Thought Prompting Elicits Reasoning."
[5] More sophisticated prompting methods can provide more accurate results. On this, see Wang, "Towards Understanding Chain-of-Thought Prompting." These are not used in this article as the intent is to replicate a non-sophisticated prompt that may be used by a law student with no experience or training in technical prompts.
[6] ChatGPT accumulated more than a million subscribers within a week of its launch: Baidoo-anu, "Education in the Era of Generative Artificial Intelligence (AI)."

## 2. The Importance of Arguments in Reasons

The contemporary common law system is based on the giving of written reasons in the determination of cases. Their importance is summed up by Summers: 'Reasons are the tools of judging, for with reasons judges resolve issues and justify decisions.'[7] As common law judges, this applies to the judges of the High Court of Australia, who are the subject of consideration in this article. Those reasons then form the basis for legal precedent, which is then used in future cases to determine issues. In this respect, it is clear that being able to understand and engage with judges' reasons is a critical component of the common law. The legal principle that follows from a particular case – described as the *ratio* or the holding – stems not merely from a particular conclusion, but from the way the conclusion was reached. It is for this reason that widely used legal databases – for example, Lexis Nexis (https://www.lexisnexis.com.au/en) or Westlaw (https://legal.thomsonreuters.com/en/westlaw) – include analyses of the judges' reasons to aid lawyers in the understanding and application of the case in the future.[8]

The reasons given by judges set out how they have reached their conclusion in a matter. In a common law system, the disposition of a particular matter before a court – for example, 'appeal allowed' – is of little substantive precedential value. The *way* that disposition is reached is of key importance.[9] Understanding what the law is – and how it may be applied in the future – requires an understanding of the reasons given in a case. In this sense, and as reasons are expressed as a persuasive piece of text – setting out the reason why a particular conclusion follows from the application of the law – they can be seen as an argument or a series of arguments.[10] For these reasons, it is a core function of a lawyer in the common law system to be able to interpret the arguments that make up those reasons. It follows that being able to identify those arguments is a skill that law students need to develop.

There are many ways in which law students can learn to interpret reasons given by the courts.[11] This could be taught explicitly in law schools, or it could be a skill implicitly taught or cultivated through learning about other aspects of the law. Basic interpretive skills can also be augmented or refined after law school while a lawyer is engaged in professional practice. Wherever this happens, there is no single necessary or definitive way to engage in the interpretive process.[12] One way – which may be familiar to many law professors – is to distil the arguments in the reasons into a formally constructed argument chain.[13] Whilst there are several forms of argument that may be relevant to law,[14] we assessed LLMs' ability to extract and reconstruct arguments into a particular form of logical argument structure – the *modus ponens* argument structure (which we explain further below) – as it is a simple and readily used form of argument.

If an LLM is capable of identifying and extracting arguments from judges' reasons in a way that accurately reflects the reconstruction of an argument by an experienced lawyer, this may facilitate a number of benefits. As one example, a benefit may be the reduction in the time spent analysing those reasons for a junior lawyer (and, as a result, the potential decrease of costs of doing so with an associated benefit to the access to justice). The benefits of an increased form of efficiency do not just relate to the practice of law, but also extend to others who need to understand the relative meaning of cases. In this respect, legal academics and law students would also benefit. It is by assessing LLMs' ability to replicate this way of achieving this essential lawyerly function that we evaluate the potential benefit that these systems may bring to the legal profession.

[7] Summers, "Two Types of Substantive Reasons"; see also Perry, "Judicial Obligation, Precedent and the Common Law."
[8] These databases now also promote an enhanced process of research using AI and GAI. For example, see Lexis+ (https://www.lexisnexis.com.au/en/products-and-services/lexis-plus) Lexis Protégé (https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-announces-new-protege-legal-ai-assistant-as-legal-industry-leads-next-phase-in-generative-ai-innovation) and Thomson Reuters' Co-Counsel (https://www.thomsonreuters.com.au/en-au/products/cocounsel.html).
[9] Perry, "Judicial Obligation, Precedent and the Common Law."
[10] Lamond, "Persuasive Authority in the Law."
[11] For an examples of a text relevant to law students that include a focus on legal reasoning and interpretation, see James, *The New Lawyer.* For a philosophical consideration of the reasons why interpretation is necessary, see Raz, "Why Interpret?"
[12] For an early consideration of the application of AI to different forms of legal arguments, see Prakken, Logical Tools for Modelling Legal Argument.
[13] Brewer, "Exemplary Reasoning."
[14] For an example of the scope and depth of the forms of legal argumentation, see, Bongiovanni, Handbook of Legal Reasoning and Argumentation; MacCormick, "Argumentation and Interpretation in Law." See also Gold, A Primer on Legal Reasoning.

## 3. Generative AI: What It is and Why It is Important

The field of AI has its roots the mid-twentieth century,[15] but a significant uptick in progress began around the turn of the twenty-first century: greater availability of data and computer processing power allowed for the training of larger neural network-based systems, heralding a 'deep learning revolution'.[16]

While early efforts in AI focused on simulating human intelligence (a research program that continues today) the term 'AI' has taken on a broader meaning, encompassing any 'machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.'[17]

These technologies have revolutionised many aspects of modern life, including developments in autonomous vehicles, facial recognition software, and the algorithms driving social media sites. A recent instalment in this story is progress in GAI, an umbrella term covering systems that can generate text, images, videos, or other data. This includes image generation systems like using generative adversarial network (GAN) (e.g. StyleGAN) and diffusion network (e.g. DALL-E) architectures. Notably it also includes Large Language Models (LLMs) – systems capable of analysing and generating text-based content.[18]

Whilst there are a number of LLMs, the best-known and first to be widely available is OpenAI's ChatGPT. The systems are trained through deep learning procedures on enormous volumes of data – for example, text corpora scraped from vast swathes of the internet.[19] Once trained, LLMs can output intelligible text and flexibly respond to a wide range of input prompts including requests to write poetry, summarise and translate text, or answer questions.[20]

The potential applications of GAI, and in particular LLMs, in the area of law are vast. Because law is a field that is fundamentally structured around various forms of legal texts – including legislation and case law – the inputs for and outputs from LLMs have much in common with materials in law. To put this another way, understanding, engaging with and interpreting the law – for practising lawyers and judges as well as students and anyone else – involves dealing with instances of words and texts. These components are the same as those with which LLMs can best interact. For this reason, there is greater potential for the beneficial use of LLMs in law than in other fields. For example, the recent study shows that ChatGPT failed to conduct IRAC analysis that was completely correct on 50 legal scenarios.[21] The ability to rapidly analyse legal texts and locate patterns, similarities or differences, or problems, has been the subject of considerable scholarly research.[22] It has already been shown that some LLMs, for example, can pass the form of exam necessary to practise law in certain jurisdictions.[23]

The ability of some LLMs to operate at this level does not mean that all LLMs can be applied in legal analysis in all jurisdictions. For some tasks, the ability to provide accurate responses relies on the availability of suitable training data – meaning that responses that are generated in relation to jurisdictions for which there is insufficient training data may not be accurate. The non-AI-based equivalent of this scenario illustrates the obvious problem: a lawyer giving advice about a legal issue in Australia, but basing that advice on precedent, legislation and other authorities predominantly from the United States, would not provide a legally relevant answer to an issue in Australia.

Another concern with the applicability of LLMs in legal contexts is their tendency to 'hallucinate'.[24] Hallucinations are instances where an LLM's output is completely, or partly, fabricated or false – without any explicit acknowledgement that this has happened. As one relevant example, the use of an LLM to draft submissions to a court resulted in the LLM inventing legal

[15] See, for example, McCarthy, "A Proposal for the Dartmouth Summer Research Project."
[16] Sejnowski, The Deep Learning Revolution.
[17] OECD definition: OECD, "AI Principles." It bears noting that the definition of AI is a matter of ongoing dispute and evolution. These subtleties of definition will not affect our argument, given that our focus in this paper on LLMs specifically.
[18] Harshvardhan, "A Comprehensive Survey and Analysis."
[19] Zhao, "A Survey of Large Language Models"; Roumeliotis, "ChatGPT and Open-AI Models."
[20] It is not uncontested that the material produced is actually new. Some see the material produced as simply being a reproduction or interpolation of material that already exists within the training data. Thus, from a certain perspective, LLMs engage in prediction rather than generation. Feuerriegel, "Generative AI."
[21] Kang, "Can ChatGPT Perform Reasoning?"
[22] See, for example, Gray, "Empirical Legal Analysis Simplified."
[23] Katz, "GPT-4 Passes the Bar Exam."
[24] Ji, "Survey of Hallucination in Natural Language Generation."

cases that were provided as authority for the claims being made.[25] Other issues could include significant safety issues associated with giving advice.[26]

Notwithstanding these issues, the potential benefits that follow from the use of GAI in legal analysis or in the provision of legal advice are considerable. For example, a 2017 OECD report listed the common barriers to accessing justice, including financial cost, time and the complexity of justice systems, a lack of legal capability and language skills.[27] Having access to legal advice is a core component of being able to have access to justice, but a substantial number of people do not seek legal advice (largely due to the considerable costs of doing so or as a consequence of overworked judicial systems).[28] In these terms, the ability to facilitate low- or no-cost (accurate) legal advice, or to increase the speed and efficiency in dealing with matters, may increase an individual's access to justice.[29] The efficiencies offered by GAI, if able to provide a form of legal advice that is accurate, may enable the costs associated with the provision of legal advice to be substantially reduced, which in turn may increase access to justice through the increased ability of individuals being able to access legal services.

For these reasons – speed, efficiency and the associated potential to increase access to justice – and in circumstances where there is considerable importance placed on the accuracy of legal analysis, as the analysis may be used to give legal advice that may impact the relevant rights or duties imposed on a person or legal entity during legal proceedings, the relative power of LLMs and GAI more broadly is not (yet) deployed widely in the legal sphere. The potential benefits need to be tempered due to the relative inability of GAI systems to perform certain aspects of legal reasoning or writing.[30] Although efficiency may follow, risks remain related to the unsupervised creation of legal arguments.[31] While there are benefits and risks of using this technology, legal services providers are promoting the use of AI and GAI extensively.[32]

Given its problematic application where accuracy and veracity are core components of the practice and operation of law, it is therefore important to determine the potential for LLMs to provide accurate responses to prompts that relate to various aspects of the provision of legal advice. This article takes a necessarily small step along this path.[33] This necessity stems from the substantial scope of what is involved in the provision of 'legal advice'. One component of this is the jurisdictional differences inherent in law. The scope of this article is intentionally limited to one component of the giving of legal advice that is relevant across common-law jurisdictions. We limit our consideration of the potential ways in which LLMs can be used to a relatively small – yet vital – sphere: the identification of arguments in judges' reasons.

The provision of reasons, as noted above, represents a vital component of the common law legal system. In limiting our assessment to the ability of LLMs to identify and analyse the arguments that are presented as part of those reasons, we are able to take one step toward using LLMs in a way that is accurate as well as useful for the wide legal community; especially in terms of the various benefits outlined earlier. In doing this, it is useful to explain in further detail the rationale behind the approach we adopt.

## 4. Rationale for Approach

In seeking to identify a suitable methodology that answers the question of whether LLMs can identify and extract arguments from judges' reasons, various methodological approaches could have been adopted.[34] For example, the field of argument mining seems relevant; however, not only do those works often adopt definitions of arguments that differ from this article, but those models also require substantial task-specific training data. As this was not available, we choose to use LLMs.[35] In selecting a methodological approach, we have sought to ensure that the second question we pose, regarding potential benefit to students, remains a principal consideration.

[25] Corcoran, "Avianca Airline Lawyer Used ChatGPT."
[26] This could include the relative absence of safeguards that allow the production of dangerous material: Wang, "Do-Not-Answer."
[27] OECD, "Equal Access to Justice for Inclusive Growth."
[28] World Justice Project, "Global Insights on Access to Justice 2019."
[29] Recent years have seen a vast number of articles proposing various ways in which AI and GAI can achieve this ultimate end. See, for example, Grossman, "The GPT Judge."
[30] Villasenor, "Generative Artificial Intelligence and the Practice of Law."
[31] Tu, "Artificial Intelligence."
[32] Thomson Reuters, "AI-Powered Legal Research"; LexisNexis, Lawyers and Robots Whitepaper.
[33] The path we have in mind may not only deliver on the potential benefits outlined earlier, but also may have the potential to fundamentally change the way law students engage with the process of education. These are, of course, purely points of speculation.
[34] Goel, "LLMs Accelerate Annotation."
[35] Lawrence, "Argument Mining."

In assessing LLMs' ability to extract arguments from judges' reasons, it is less interesting to question whether LLMs can provide *an* answer when prompted in these terms – as, in the most pedantic sense, even a nonsensical answer would require this question to be answered 'yes'. It is more interesting to assess how accurately the systems' outputs represent the relevant arguments: do they, for example, provide argument reconstructions that are completely nonsensical, minimally incorrect or correct?

To identify the relative accuracy of the answer provided, as set out in the methodology section below, we use a rubric to compare the LLMs' argument reconstructions with arguments reconstructed by an experienced lawyer. In order to add specificity to the outcome generated by the LLMs and to enable a comparison to be made with the argument identified in the reasons by an experienced lawyer, we ask the LLMs to reconstruct arguments in a *modus ponens* argument form – that is arguments with the form:

*If A then B*
*A*
*Therefore, B*

In other words, we ask both the human and the LLMs to reformulate the arguments in the reasons into a chain of if/then arguments that result in the final disposition in the case. (We expand on this argument form in Section 5.) By doing this, the aim is not to only identify instances where arguments are explicitly presented in if/then form by the judges; the aim is to identify *all* the arguments (regardless of the form in which they are presented by the judge/s) and to reconstruct them into a single or chain of arguments that follow the if/then form.

As stated earlier, while there are a number of ways that arguments can be made,[36] the if/then form is relatively simple and frequently invoked.[37] For these reasons, the formulation or reformulation of the arguments within the judges' given reasons into if/then arguments has the advantage of less reconstruction in comparison to creating another form of argument. Furthermore, given the relevance of the if/then approach to the construction of legal arguments, having an ability to distil arguments in this way also has some relevance to at least one skillset that law students may seek to develop.

We adopted a broadly student-focused approach in designing the methodological approach outlined below. The aim was to use the LLMs in a way that stood a plausible chance of benefiting an average law student, where the average law student is one who does not have expertise in either logic, argument construction, or information technology/coding. For this reason, that we did not use a complex or technical form of prompting (even where a more technical form of prompt may have yielded a more accurate output).[38] Similarly, when testing the LLMs, we did not engage them through an API, instead using the same form of prompt that anyone with an internet connection and a basic understanding of the way LLMs can be used could do.[39]

A similar rationale was applied to the source of the materials that we explicitly provided to the LLMs. We limited our consideration of judges' reasons to those published by the High Court of Australia that are also freely available on the internet. We did not, for example, utilise material – available behind any sort of paywall – that may have provided a commentary or additional insight into those cases specifically. This had a dual benefit. First, it avoided any licensing issues associated with uploading proprietary content to the LLMs. Second, by using material that was publicly available, we ensured that the same conclusions or outputs could be derived by anyone – without the need to specifically gain access to premium materials. Even though many law students may already have access to those materials, most members of the public do not.

The specific focus on apex court decisions also has a twofold rationale. The first aspect is due to the wide availability of the reasons. The second is because apex court judges – because of the relative importance of their decisions and because of the common ability to restrict their own caseload through grants of special leave – are afforded sufficient time to write their reasons.[40]

[36] For some examples of just the deductive forms of argument, see Gold, A Primer on Legal Reasoning, 65–84.
[37] The frequency of their use is noted in Walton, "Are Some Modus Ponens Arguments Deductively Invalid?" Consider also the arguments extracted in the course of this study; a number of if/then arguments are evident in even small sections of the judges' reasons.
[38] Baidoo-anu, "Education in the Era of Generative Artificial Intelligence (AI)," 52–62.
[39] We did not consider the paid forms of ChatGPT and Claude to be beyond the means of most law students – who are already committing substantial sums to the costs of law school.
[40] This is not to say that the judges in apex courts are not under considerable time pressures – merely that their role requires the reasons to be properly considered.

In attempting to answer the research questions, we sought to ensure the focus remained on the LLMs' ability to identify arguments – and not, for example, to be able to discern the most legally relevant set of reasons in any given text. LLMs' ability to identify arguments has been a hot research topic in the field of natural language processing. However, the definition and form of the arguments that have been explored in those works are different from the legal argument forms that we seek to extract.[41] In this regard, the process in which we engage should be considered a new task rather than a conventional NLP argument extraction task.

The cases selected in our task include five judges providing reasons as a unanimous decision, or as a minority/majority split. To avoid spurious answers, the prompts used in seeking an output from the LLMs specified the judge or judges' reasons that should be considered. The request always specified the judges by name in either the unanimous or majority decision. We did not, for example, request the arguments with the most significant legal relevance, as the aim was to explore whether the LLM could correctly identify arguments and not whether it could discern or identify a majority decision.

## 5. Methodology

To assess whether LLMs can accurately extract legal arguments, we treated LLM outputs akin to student outputs – that is, we sought to assess the LLM's output's accuracy in comparison with a predetermined form of answer and marking criteria. This necessitated both the creation of a marking rubric that included specific marking criteria and a rationale for the points allocation to be used by a marker as well as the identification of the form of the argument that was actually included in the reasons. In this section, we outline the methodological approach associated with these steps and further explain additional aspects of the methodology associated with the form of the argument and the cases used.

The creation marking rubric and the identification of the form of the argument contained within the reasons were carried out by a lawyer/legal academic with several years' practice and teaching experience across several jurisdictions and by a philosopher with expertise in, and experience teaching, the analysis and formal reconstruction of arguments. The same two individuals were also responsible for assessing the accuracy of the LLMs' outputs. This was conducted after a moderation exercise – relating to a different case (*Ismail*, detailed below) – to ensure the usability of the rubric against an LLM output.

It is a common practice in NLP to use domain experts to evaluate LLM outputs. In legal NLP, the cost of human evaluation is high. This has been acknowledged as a bottleneck or limitation in the annotation or assessment of outputs in this area.[42] Those same factors impacted the present study. (This represents a limitation in terms of the study's conclusions.)

With regard to these factors, it is not uncommon to engage a small number of qualified legal experts to conduct human evaluation.[43] Further, it is also not uncommon to use only two human experts.[44] To maximise the effectiveness of the assessors used, we recruited assessors with relevant and complementary expertise: a philosopher with expertise in, and experience teaching, the analysis and formal reconstruction of arguments (to assess the LLMs' ability to faithfully and clearly capture the logical structure of the arguments); and, a lawyer/legal academic to assess the LLMs' ability to preserve the subtleties of the legal terminology contained within the extracted arguments.

A generic rubric was created so it could be used to mark outputs related to any of the cases. (The rubric is included as Appendix A at https://github.com/qulizhen/LTHJ_legal_argument_identification) It is a form familiar to many in higher education. Out of a total of 20 marks, the rubric assessed (and allocated a mark total in relation to) the ability of the LLMs to:

1. identify the disposition (3 marks)
2. identify the premises and conclusions leading to the final conclusion (3 marks)
3. identify the premises and conclusions in prior arguments (6 marks)
4. locate arguments in the text (with a paragraph number/s) (3 marks)
5. use a *modus ponens* argument structure to reformulate the arguments (5 marks).

[41] Mok, "LLM-Based Frameworks"; Rescala, "Can Language Models Recognize Convincing Arguments?"; Chen, "Exploring the Potential of Large Language Models."
[42] Savelka, "Can GPT-4 Support Analysis of Textual Data?"
[43] See, for example, Kang, "Can ChatGPT Perform Reasoning?"
[44] See, for example, Park and Cardie, "Identifying Appropriate Support for Propositions."

In assessing the ability to identify premises, conclusions and underlying logical form of argument, the rubric targeted a grasp of the core structural components of arguments.[45] Then, in assessing the ability to locate the relevant elements in the text, the rubric targeted the ability to map the reconstructed argument back onto the source document (this is useful for markers, and also for any would-be student end-users). The mark relevant to each criterion was allocated based on a scheme that described tiers of criterion satisfaction. This was set out prescriptively in the rubric to facilitate consistency across the markers. For example, in relation to category 4, 'If a precise *modus ponens* format is adopted across the whole argument chain, 5 marks should be awarded.' The various terms used in each criterion were also defined and described within the rubric.

The court cases that were used in the study were the first six published decisions of 2024 from the High Court of Australia: *Harvey v Minister for Primary Industry and Resources* ('Harvey'); *Lesianawai v Minister for Immigration, Citizenship, and Multicultural Affairs* ('Lesianawai'); *The King v Rohan (a pseudonym)* ('Rohan'); *Carmichael Rail Network Pty Ltd v BBC Chartering Carriers GmbH &Co KG* ('Carmichael'); *Xerri v The King* ('Xerri'); and *Ismail v Minister for Immigration, Citizenship and Multicultural Affairs* ('Ismail').[46]

The first five cases noted immediately above – we refer to these as 'the cases' – were the cases used to obtain LLM outputs that were marked and that make up the results noted in the section below. The sixth case – *Ismail* – was randomly selected from the cohort of six cases to be used as a test case for the markers to perform a moderation exercise with the rubric.

While small, the sample of cases was diverse. The cases represented a range of different forms of legal issues – including native title, criminal law, statutory interpretation and immigration – and different kinds of matter – both appeals and applications in the Court's original jurisdiction. The cases were also different in their total extent – with reasons that made up the relevant texts ranging from 23–53 pages.

The PDF reasons provided by the judges in the cases were examined by an experienced human lawyer to identify the form of the argument that was provided by the majority in each set of reasons. Marker 1 reproduced the form of the arguments into a *modus ponens* form of formal argument chain and included the paragraph numbers for the premises and conclusions that were identified. Also, as it is infrequently the case that a set of reasons presents only one argument, chains of argument – where the conclusion to one argument represents one of the premises to a subsequent argument – were identified.

Once the argument chain within the majority reasons was identified, it was reviewed by Marker 2. As a result of the review, a final agreed chain of arguments was then created as a 'sample answer' for each of the cases. (The sample answers are included as Appendix B at https://github.com/qulizhen/LTHJ_legal_argument_identification)

Using the PDF versions of the cases as an uploaded attachment accompanying the prompt, the LLMs were then asked to identify the argument chains within the cases. A single prompt was created to do this. (The prompt is included in Appendix C at https://github.com/qulizhen/LTHJ_legal_argument_identification) The prompt used is compendious. It was created as a result of a need to create a non-technical form of prompt to achieve the desired output required by the rubric. While it is possible to use sophisticated prompting techniques that move away from standard single prompts in basic written English,[47] the aim was to use the sort of prompt that a law student with no other IT/coding/prompting experience may be able to formulate.

Two versions of two different LLMs were assessed: ChatGPT4o, ChatGPT4, Claude 3.5 and Claude 3.0. Using the prompt, responses were sought from each model for each of the cases, yielding 20 responses in total. Across both systems and all the versions, the prompts were entered on the same day, with each prompt commencing a new dialogue consisting of *only* this prompt. No further prompting or clarifications were entered. The first response following the prompt was recorded as the given answer relative to each of the cases and each of the systems/versions. (The outputs are included as Appendix D at https://github.com/qulizhen/LTHJ_legal_argument_identification)

[45] Dutihl Novaes, "Argument and Argumentation.

[46] *Harvey v Minister for Primary Industry and Resources* [2024] HCA 1, 7 February 2024; *Lesianawai v Minister for Immigration, Citizenship, and Multicultural Affairs* [2024] HCA 6, 6 March 2024; *The King v Rohan* [2024] HCA 3, 14 February 2024; *Carmichael Rail Network Pty Ltd v BBC Chartering Carriers GmbH &Co KG* [2024] HCA 4, 14 February 2024; *Xerri v The King* [2024] HCA 5, 6 March 2024; and *Ismail v Minister for Immigration, Citizenship and Multicultural Affairs* [2024] HCA 2, 7 February 2024.

[47] Paweł Korzyński, Grzegorz Mazurek, Pamela Krzypkowska, and Artur Kurasiński, "Artificial Intelligence Prompt Engineering as a New Digital Competence: Analysis of Generative AI Technologies Such as ChatGPT," *Entrepreneurial Business and Economics Review* 3 (2023).

The sample answer and the rubric were then used to assess the answers given. This process involved the separate marking of each output by human assessors with experience in the identification of arguments.[48] The assessors were the same lawyer and philosopher who had created the rubric.[49] Prior to the marking commencing and prior to the markers being given the outputs, the various LLMs' outputs were anonymised and provided with a random numerical identifier. This was done to provide a form of blind marking and to avoid bias being introduced into the marks given to the outputs. After marking the outputs in accordance with the rubric, the marks were recorded for each of the outputs. (The outputs, after being de-anonymised, are included as Appendix E at https://github.com/qulizhen/LTHJ_legal_argument_identification).

## 6. Results

For anyone unfamiliar with the relative ability of LLMs to produce responses to queries, the results are startling. When prompted, within seconds each of the LLM versions provided a detailed response that – superficially, at least – approximated the arguments presented in the judgments. In this respect, the responses provided at least approximated a form of answer to the prompt provided. However, for anyone familiar with LLMs, it will come as no surprise that the systems often presented inaccurate responses in a very confident way.

In the following subsections, we consider some of the ways in which the outputs vary. We also provide just one illustration of the nature of the differences between the system outputs – in effect, demonstrating a good and a bad example of the outputs. We also include some brief observations regarding differences in markers' marks before, in a final very brief subsection, summarising the differences.

### *6.1 Differences in Marks Across Systems and Versions*

On this basis, at least at the most basic binary level, each of the LLMs tested *can* respond to prompts requesting that the arguments in the judges' reasons be identified and extracted. The relative accuracy of the outputs, however, varies considerably. When considered in tabular form ranked from the highest marked output to the lowest, as shown in Table 1, the relative difference in the systems' ability to accurately identify the argument chains becomes apparent.

[48] For a similar approach, see Kang et al., "Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a Lawyer?" See also earlier notes in the article that note the inherent problem with locating multiple domain experts and the similarity of approach with other low-number assessor projects.

[49] Whilst this inherently leads to a potential bias, the potential of this was reduced by: creating the rubric prior to obtaining the outputs; randomly selecting the subject cases; anonymising the output identity (in terms of the system that created the output); and, ensuring the assessors assessed the outputs independently.

**Table 1. LLM outputs ranked in order of average marks**

| **Average mark (of two markers) out of 20** | **System – version** | **Case** |
|---|---|---|
| 18 | Claude – 3.5 Sonnet | *Xerri* |
| 17 | Claude – 3.5 Sonnet | *Harvey* |
| 17 | Claude – 3.5 Sonnet | *Carmichael* |
| 15 | Claude – 3.5 Sonnet | *Rohan* |
| 14 | Claude – 3.5 Sonnet | *Lesianawai* |
| 12 | Claude – 3 Opus | *Xerri* |
| 12 | Claude – 3 Opus | *Lesianawai* |
| 11 | Claude – 3 Opus | *Rohan* |
| 10 | Claude – 3 Opus | *Harvey* |
| 10 | GPT – 4o | *Lesianawai* |
| 10 | GPT – 4 | *Lesianawai* |
| 10 | GPT – 4o | *Harvey* |
| 10 | GPT – 4 | *Harvey* |
| 9 | GPT – 4 | *Carmichael* |
| 9 | GPT – 4o | *Xerri* |
| 8 | GPT – 4 | *Xerri* |
| 8 | Claude – 3 Opus | *Carmichael* |
| 6 | GPT – 4o | *Rohan* |
| 5 | GPT – 4o | *Carmichael* |
| 4 | GPT – 4 | *Rohan* |

The range of marks across Table 1 is starkly apparent – ranging between 4/20 and 18/20. Within this study, it is not possible to place these marks in the context of (human) students' marks for the same task, as students were not given the same task. However, based on the highest and lowest outputs, and on previous experience grading exercises such as this in university-level students, the highest mark would be what would be expected to be seen in an excellent law student (at the highest available level of a law student). The output that was attributed the lowest mark would likely be much closer to a fail or pass grade for the same student.

When the mark allocated to each of the outputs is considered, several key findings become apparent. The first and most stark of these is that when the average mark of the two markers is considered and the LLM marks are ranked, Claude generally achieves higher marks than ChatGPT. The five Claude 3.5 outputs achieved the highest average five marks – from 14/20–18/20; the next three highest marks were Claude 3.0 – from 11/20–12/20. Another Claude 3.0 output achieved 10/20 jointly with several ChatGPT outputs. The relative difference between the two systems' abilities to identify arguments is also apparent from the average grade awarded to each system (as an average of the average of the two markers' marks – what we will call the 'system average'): the system average mark for Claude (3.0 or 3.5) output is 13.4/20; the system average mark for ChatGPT (4 or 4o) output is 8.1/20. This is starker when the individual versions are considered: ChatGPT 4 = 8.2/20; Claude 3.0 = 10.6/20; ChatGPT 4o = 8/20; Claude 3.5 = 16.2/20.

In continuing to consider the system average, the highest mark attributed to any of the systems/versions is 16.2/20 (Claude 3.5). As the lowest system average attributed to any of the systems/versions is 8/20 (ChatGPT 4o), it is clear that the best-performing version – Claude 3.5 – achieves more than double the marks of the worst-performing version – ChatGPT 4o. It can also be seen that not only do both versions of Claude perform, on average, better than GPT's versions, but also that the more advanced version of ChatGPT (GPT4o) performs worse – albeit only marginally – than the less advanced version (GPT4) with averages of 8/20 and 8.2/20 respectively.

In contrast, Claude 3.0's and Claude 3.5's system average marks are 10.6/20 and 16.2/20, respectively. In contrast, the Claude outputs score higher than the average mark with only two outputs from Claude 3.0 scoring below the average; all of the Claude 3.5 outputs exceed the average, with the lowest average mark for that version being 14/20. When considered against the average mark across the entire cohort – of 10.75/20 – it is only Claude 3.5's outputs that exceed that mark – as Claude 3.5's outputs scored on average 16.2/20. The other versions fell either marginally or considerably below the overall cohort average.

What is also apparent is that while the Claude (3.5 and 3.0) outputs cluster according to the version of the system – with 3.5 outputs all scoring higher than 3.0 outputs – this is not the case with the ChatGPT (4o and 4) outputs. The Claude outputs seem to reflect the level of sophistication of the versions – with Claude 3.5 being the newer iteration. There is a marked difference between the two versions' ability to accurately identify and extract arguments.

This is not the case with ChatGPT. There is relatively little difference between the marks for the two versions of ChatGPT for several cases; the ChatGPT outputs pair up with similar marks being given to the outputs related to each case. The outputs for *Lesianawai* and *Harvey* from both ChatGPT versions get the same mark. For *Xerri*, they are only one mark apart; for *Rohan* they are two marks apart. For the final case, *Carmichael*, they are three marks apart. This suggests that, in contrast to Claude, there is little difference between the two ChatGPT systems' ability to extract arguments.

When considered in terms of the individual legal case towards which the LLMs were pointed, a clear homogeneity can be seen across most of the cases. The overall average mark (as an average of the two markers' marks across the whole cohort) was 10.75/20. When the average mark for each of the cases is considered (as an average of the two markers marks from all systems/versions for a single case), these did not differ substantially from the overall average. The case averages were: *Harvey* 11.75/20; *Xerri* 11.75/20; *Lesianawai* 11.5/20; *Carmichael* 9.75/20; and *Rohan* 9/20. For all but *Rohan*, the average for each case was within one mark (out of 20) of the overall average; for *Rohan*, it was 1.75 marks below the cohort average. This demonstrates that there was relatively consistent uniformity across the marks for each of the cases – which indicates that the relative difference in legal topic of the cases was relatively inconsequential to the LLMs' ability to identify and extract arguments.

The same could also be said in terms of the page extent of each of the cases/reasons provided to the LLMs. When each case is represented as an [average mark]/[page extent] format, the results are: *Harvey* 11.75/33; *Xerri* 11.75/14; *Lesianawai* 11.5/13; *Carmichael* 9.75/26; and *Rohan* 9/15. When considered in this way, there is no apparent correlation between the page extent of the reasons and the LLMs' ability to extract arguments accurately. In other words, the length of the documents seemed to have no impact on the ability of the systems to accurately identify and extract arguments.

Whilst the LLMs were directed specifically towards the majority reasons via the prompt, the same conclusions as expressed immediately above can be drawn from a consideration of the form of the reasons. The reasons represented different expressions of the legal disposition. Two cases (*Carmichael* and *Lesianawai*) were unanimous – with the former expressed as a single set of reasons and the latter being a single judge's reasons with which all of the other four judges expressed agreement); *Rohan* and *Xerri* were majorities of three judges; and *Harvey* was a majority of four judges. There is no apparent correlation between the way the decisions are expressed and the average mark. However, as noted, where the prompt specified the reasons from which the arguments should be extracted, this is unsurprising.

Similarly, and equally unsurprisingly, the fact that four of the matters were in the Court's appellate jurisdiction and one (*Lesianawai*) was in its original jurisdiction does not appear to have fundamentally impacted the LLMs' ability to accurately identify the arguments in the reasons – as *Lesianawai* was ranked third in terms of the average mark across the systems.

### *6.2 Specific Failures in the Systems – An Example*

The assessment category that was least well answered by the LLMs was the category that required paragraph numbers in which the arguments given in the reasons were located. This was the case across both Marker 1's and Marker 2's assessments of the appropriate marks – with 1.3/5 and 1.65/5, respectively, being awarded by each marker. From an assessor's perspective, this was relatively surprising – as the citation of paragraphs (once the premises and conclusions of the argument had been identified, often correctly) is a straightforward task. Instead, it was often the case that – despite the prompt including a specific requirement to include paragraph numbers – these were often omitted.

The systems' lower marks in the categories that relate to paragraph numbers may reflect the fact that ChatGPT and Claude as Large *Language* Models are trained on and deal with language and not numbers. LLMs like ChatGPT have been shown to be

poor at performing mathematical calculations.[50] However, the nature of this study does not allow any conclusions to be drawn other than the mere observation of the correlation of the lowest marks with the only assessment category that involves numerical values.[51]

In more specific terms relating to the task at hand, the LLMs performed reasonably well at reconstructing the arguments in accordance with a *modus ponens* form, with the average mark in this category being 3/5. This is perhaps unsurprising given the formulaic nature of this language-based task. However, it is noteworthy that in several instances systems failed on this front, deviating substantially from basic *modus ponens* structure. The following is an example of such a failure (taken from GPT-4o's reconstruction of the *Xerri* case):

> 1. **Premise 1 (P1):** If an act increases the penalty for an offence, the increased penalty only applies to offences committed after the commencement of that act.
>
> - **Reference:** Section 19(1) of the Crimes (Sentencing Procedure) Act 1999 (NSW) (the Procedure Act) 【7:4†source】.
>
> 2. **Premise 2 (P2):** The appellant's offending occurred prior to the commencement of the current s 66EA of the Crimes Act.
> - **Reference:** Paragraph 41 【7:4†source】.
>
> 3. **Conclusion 1 (C1):** Therefore, if the current s 66EA of the Crimes Act increases the penalty for the same offence, the increased penalty should not apply to the appellant's case.

This can be contrasted with the sample answer for the same case:

> IF the replacement of s 66EA constituted an increase in the penalty for the "offence" already created by former s 66EA, THEN the current offence is different to the old, and is a new offence. [41]
>
> the replacement of s 66EA constituted an increase in the penalty for the "offence" already created by former s 66EA [41], [65]
>
> Therefore, the current offence is different to the old, and is a new offence [41], [65], [66]

When compared with the sample answer, the ChatGPT 4o output's relative absence of a clear *modus ponens* format – If P, then Q; P; therefore, Q – can be seen. (What can also be seen is a deficiency in the paragraph number citations.) These relative failings can be further contrasted with Claude 3.5 Sonnet's reconstruction of *Xerri*:

> 1. If the factual ingredients or elements of former s 66EA differ significantly from current s 66EA, then current s 66EA is a new offence. (Implied premise)
>
> The factual ingredients or elements of former s 66EA differ significantly from current s 66EA. (Para 65)
>
> Therefore, current s 66EA is a new offence. (Para 65)

As is apparent, Claude 3.5's reconstruction is, to all intents and purposes, identical to the sample answer in terms of the argument construction and content. The only substantive difference does not relate to the form of the argument; as the Claude 3.5 output suggests one of the premises is implied – whereas the sample answer attributes this to a specific paragraph.

When considering these forms of output, not only does the ability of Claude 3.5 to reconstruct an argument in the same terms as the sample answer (that was created by an experienced lawyer) become apparent and clear, but the failings in the output from ChatGPT 4o also become apparent. This represents a key – but not the only – stark illustration of the differences between the two systems. In these terms, it is clearly apparent that there are very significant differences between the two systems – even between the two systems' flagship versions (of ChatGPT 4o and Claude 3.5)

[50] Frieder, "Mathematical Capabilities of ChatGPT."
[51] It was noted by one of the reviewers, and we agree, that the citation of paragraph numbers may well have been improved through a simple follow-up prompt. (That would, however, go beyond the one-shot strategy that we adopt.)

### *6.3 Marker Differences*

The results also revealed a difference in the way that Markers 1 and 2 marked the LLM outputs – despite both marking the submissions in accordance with the same detailed rubric. The average difference between the two markers was 3.05/20. At times, the difference between the two was as big as 8/20. Typically, Marker 1 graded more harshly – as in only two out of the 20 outputs marked did Marker 1 give a higher mark than Marker 2. These differences may be due to differences in the marker's disciplinary backgrounds, which perhaps led to different appraisals of the salience of certain errors. (For example, Marker 1 took a much narrower view of what constituted the appropriate formulation of the conclusion relative to Marker 2.) However, our data alone are insufficient to establish this. It may also be that assessing how successfully an answer extracts an argument from the source text involves some irreducible subjectivity, even with a detailed rubric. In any event, there was a moderately strong correlation between the overall rankings of each output resulting from the two markers' blinded and independent marking, leaving the broad conclusions drawn in this article intact.

### *6.4 Summary of Results*

Even on this basic form of experimentation, the most obvious result is that the choice of LLM matters when answering the question of whether LLMs can identify arguments. While all the systems/versions were able to identify arguments and argument chains to some degree, some performed far better when measured against the assessment rubric.

The relative difference between systems is stark. As noted, while the assessment scheme adopted is not compared directly to (human) student outputs, the relative standard exhibited from the top mark – 18/20 for Claude 3.5 – and the lowest mark – 4/20 for ChatGPT 4 – is anecdotally the difference between the output expected of a student who would achieve the highest available grade and a student who may achieve a fail or very bare pass grade.

The starkness of the differences, coupled with the clear separation between the ability of the two systems – ChatGPT and Claude – to perform the requested task is also obvious. The far more well-known system – ChatGPT – performed, almost universally, worse than both Claude versions. While there was a clear difference in marks between the two versions of Claude – with the premium 3.5 performing uniformly better in the assessments – the same cannot be said for ChatGPT; the marks obtained by the premium ChatGPT 4o version were frequently indistinguishable from ChatGPT 4.

In general, both systems and their versions performed worst in relation to the requirement to locate paragraph references. Within the marking process, this could be attributed to a variety of reasons. This includes both the inaccurate attribution of paragraph references as well as the omission of paragraph references altogether. In this respect, the problems with the paragraph numbers represented the most obvious 'blind spot' across the marking rubric.

In consideration of some of these key results, it is now possible to provide answers to the research questions posed within this article.

## 7. Research Question Answers

In the two subsections that follow, we provide brief responses to the research questions posed in this article before providing our conclusions (including our recommendations for further research).

### *7.1 Do the LLM Outputs Accurately Identify and Extract the Legal Arguments?*

The briefly stated, yet necessarily hedged, answer to this question is: yes … sometimes. LLMs can identify and extract legal arguments from court cases to varying degrees of accuracy. Given this result, there is substantial *potential* benefit for students; however, within the terms of the one-shot approach we adopt, this requires that students use the sort of system that can provide the most benefit.

The results show that LLMs can extract legal arguments from court judgments. However, the level at which the extraction can be considered correct or accurate varies considerably across model types and specific model versions. In this regard, when seeking to answer the question it is necessary to consider the outputs distinctly – at least on the basis of the two different systems.

Looking to the best-case scenario that is reflected in the outputs provided by Claude 3.5, the answer to this research question is an unequivocal 'yes'. The marks for Claude 3.5 range from 15/20–18/20. Marks of this sort would be of the kind achieved

by some of the best law students assessed in similar tasks. In this regard, and as was seen in the Claude 3.5 example extracted earlier, the GAI outputs do accurately identify and extract the legal arguments from within the cases.

The same cannot be said for the outputs from ChatGPT. Whilst the highest ChatGPT mark was 10/20 (achieved by both ChatGPT 4 and ChatGPT 4o), six of the 10 Chat GPT marks were 9/20 or below (with the lowest mark being 4/10). The outputs provided by ChatGPT – regardless of the version – demonstrate that while there were arguments extracted from the reasons, there were significant inaccuracies in several respects.

Claude 3.0 represents something of a middle ground between the two ChatGPT versions and Claude 3.5. Four of the five outputs range between 12/20 and 10/20. The other output was marked 8/20. This places four of Claude 3.0's outputs at or above the highest scoring ChatGPT outputs.

These three different classifications of the outputs clearly illustrate that it is possible for LLMs to accurately identify and extract arguments from legal reasons; however, they also illustrate that not all LLMs can do this accurately and, in fact, some do a very poor job compared with others.

### *7.2 Are LLM Outputs Beneficial for Law Students?*

Given the qualified findings above, the answer to this question must also be qualified. The question of benefit could be viewed in different ways – both pedagogically and practically. The aim here is not to cover all possible benefits. Instead, we separate the idea of potential benefit into two separate ideas. The first is whether the LLMs' outputs could benefit students in getting better grades; the second is whether LLMs may benefit students more generally.

In considering the first, it seems clear that the use of Claude 3.5 would – for most students – benefit students in getting a better grade. The highly accurate identification and extraction of arguments that Claude 3.5 can achieve represents that which would be seen only in the very best students. It therefore stands to reason that all other students would, therefore, benefit (in terms of their grades) from using a system like Claude 3.5.

It is equally clear that the use of a system like ChatGPT would likely not benefit most students. The identification and extraction of the sorts of arguments that were output by ChatGPT would fall below that expected for most students.

The second idea is the way in which LLMs may benefit students more generally. It should not be ignored that either of the LLMs *could* be used as a pedagogical/learning aid in order to engage or stimulate further learning; a keen student could use the outputs to test their own understanding or knowledge. This may assist them both in terms of their grades as well as in the enhancement of their knowledge more generally. However, it is also conceivable that students could use one of these systems as a way of *avoiding* having to learn how to extract the arguments themselves. The existence of a shortcut tool like this is not unprecedented in education. (The introduction of the pocket calculator springs to mind.) The potential for the avoidance of learning this kind of skill to have an impact on law students' ability to be effective lawyers does, however, appear to be relevant. If nothing else, where there is a considerable difference in the ways different systems can accurately identify and extract arguments, there is a need for some level of knowledge in order to critically assess what output is being provided. Thus, if students uncritically use current versions of ChatGPT (the most well-known LLM) to extract legal arguments, not only might their outputs be poor, but the negative impact on their legal argument reconstruction skills may be compounded. Conversely, if students (perhaps with the guidance of educators) take a more critical approach, and actively assess the accuracy of LLMs' argument reconstructions for themselves, these tools could be used to further hone argument reconstruction skills. This discussion highlights that a delicate pedagogical approach is required to successfully incorporate generative AI tools into legal education (as has been discussed elsewhere).[52]

## 8. Conclusion

The answers that have been provided above are hedged: it is *sometimes* possible for LLMs to accurately identify and extract argument chains from judges' reasons and it may *sometimes* be beneficial for law students to use LLMs to do this. This hedging is necessary, in part, due to the substantially different abilities of the two LLMs to identify and accurately extract the arguments in judges' reasons, and in part due to the complex factors affecting how the use of AI tools translates into student learning outcomes. (The finding should also be understood in terms of the previously acknowledged limitation due to the limited number of assessors.)

[52] Koplin, "Tailoring University Assessment."

It is this finding – that LLMs vary so considerably in their ability to accurately extract arguments – that represents the key finding of this study. Being aware of this enables legal educators and lawyers to remain focused on the need for critical awareness regarding the human skill of identifying and extracting arguments. By retaining this focus, and by ensuring that we do not naively consider all LLMs to be 'good enough', these systems can be seen to provide benefit in different ways.

While the need for specific skills-based education of this sort goes beyond the application of a hunch or a tacit intuition, it does demonstrate that there is – at least as relates to the state of generative AI technology today – an ongoing need to retain the human-centric focus on the attainment of the lawyer's skillset in the process of education.

## Bibliography


Baidoo-anu, D, and L Owusu Ansah. "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning." *Journal of AI* 7, no 1 (2023): 52–62. https://doi.org/10.61969/jai.1337500.

Bongiovanni, Giorgio, Gerald Postema, Antonino Rotolo, Giovanni Sartor, Chiara Valentini, and Douglas Walton (eds), *Handbook of Legal Reasoning and Argumentation*. Dordrecht: Springer, 2018.

Brewer, Scott. "Exemplary Reasoning: Semantics, Pragmatics, and the Rational Force of Legal Argument by Analogy." *Harvard Law Review* 109, no 5 (1996): 923–1028. https://doi.org/10.2307/1342258.

Carlile, Winston, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. "Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Iryna Gurevych and Yusuke Miyao, 621–31. Melbourne: Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/P18-1058.

Chen, Guizhen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. "Exploring the Potential of Large Language Models in Computational Argumentation." *arXiv*, July 1, 2024. https://doi.org/10.48550/arXiv.2311.09022.

Corcoran, Kieran. "Avianca Airline Lawyer Used ChatGPT and Cited Fake Cases, Judge Says." *New York Times*, May 27, 2023. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

Dutilh Novaes, Catarina. "Argument and Argumentation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Feuerriegel. Stanford, CA: Stanford University Press. https://plato.stanford.edu/entries/argument.

Feuerriegel, Stefan, Jochen Hartmann, Christian Janiesch, and Patrick Zshech (eds). "Generative AI," *Business & Information Systems Engineering* 66 (2024): 111–126. https://doi.org/10.1007/s12599-023-00834-7.

Frieder, Simon, Luca Pinchetti, Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. "Mathematical Capabilities of ChatGPT." In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*. 2023.

Goel, Akshay, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, et al. "LLMs Accelerate Annotation for Medical Information Extraction." *arXiv*, December 4, 2023. https://doi.org/10.48550/arXiv.2312.02296.

Gold, Michael Evan. *A Primer on Legal Reasoning*. Ithaca, NY: Cornell University Press, 2018.

Gray, Morgan A., Jaromir Savelka, Wesley M. Oliver, and Kevin D. Ashley. "Empirical Legal Analysis Simplified: Reducing Complexity Through Automatic Identification and Evaluation of Legally Relevant Factors." *Philosophical Transactions of the Royal Society A*. 2024. https://doi.org/10.1098/rsta.2023.0155.

Grossman, Maura R, Hon Paul W Grimm, Daniel G Brown, and Molly Xu. "The GPT Judge: Justice in a Generative AI World." *Duke Law & Technology Review* 23, no 1 (2023). https://complexdiscovery.com/wp-content/uploads/2023/05/The-GPTJudge-Justice-in-a-Generative-AI-World-Authors-Copy.pdf.

Guha, Neel, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, et al. "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." *arXiv*, August 20, 2023. https://arxiv.org/abs/2308.11462v1.

Harshvardhan, GM, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. "A Comprehensive Survey and Analysis of Generative Models in Machine Learning." *Computer Science Review* 38 (2020): 100285. https://doi.org/10.1016/j.cosrev.2020.100285.

Holland, James A., and Julian Webb. *Learning Legal Rules*, 8th ed. Oxford: Oxford University Press, 2013.

James, Nickolas, Rachael Field, and Jackson Walkden-Brown. *The New Lawyer,* 3rd ed. Chichester: Wiley, 2023.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55, no 12 (2023). https://doi.org/10.1145/3571730.

Kang, Xiaoxi, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. "Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a Lawyer?" In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13900–923, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.929.

Katz, Daniel Martin, Michael James Bommarito, Shang Gao, and Pablo Arredondo. "GPT-4 Passes the Bar Exam." *Philosophical Transactions of the Royal Society A* 382 no *2270* (2024). https://doi.org/10.1098/rsta.2023.0254.

Koplin, J., R. Sparrow, J. Hatherly, N. Rivers, and I. Williams. "Tailoring University Assessment in the Age of ChatGPT." *Monash Lens*, May 15, 2023. https://lens.monash.edu/@politics-society/2023/05/15/1385696/tailoring-university-assessment-in-the-age-of-chatgpt.

Korzyński, Paweł, Grzegorz Mazurek, Pamela Krzypkowska, and Artur Kurasiński. "Artificial Intelligence Prompt Engineering as a New Digital Competence: Analysis of Generative AI Technologies Such as ChatGPT." *Entrepreneurial Business and Economics Review*, 3 (2023).

Lamond, Grant and The Harvard Review of Philosophy. "Persuasive Authority in the Law:" *The Harvard Review of Philosophy* 17, no 1 (2010): 16–35. https://doi.org/10.5840/harvardreview20101712.

Lawrence, John, and Chris Reed. "Argument Mining: A Survey." *Computational Linguistics* 45, no 4 (2019): 765–818. https://doi.org/10.1162/coli_a_00364.

LexisNexis. *Lawyers and Robots Whitepaper*, 2017. https://www.lexisnexis.com.au/__data/assets/pdf_file/0003/187644/Lawyers_and_Robots_Whitepaper.pdf.

MacCormick, Neil "Argumentation and Interpretation in Law". *Ratio Juris* 6, no 1 (1993): 16–29. https://doi.org/10.1111/j.1467-9337.1993.tb00135.x.

Makridakis, S. "The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms." *Futures* 90 (2017): 46.

J. McCarthy, M. Minsky, N. Rochester, and C. E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." *AI Magazine* 27, no 4 (1955). https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904.

Mok, Jisoo, Mohammad Kachuee, Shuyang Dai, Shayan Ray, Tara Taghavi, and Sungroh Yoon. "LLM-Based Frameworks for API Argument Filling in Task-Oriented Conversational Systems." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, edited by Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, 419–26. Mexico City: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.naacl-industry.36.

OECD. "Equal Access to Justice for Inclusive Growth." March 27, 2019. https://www.oecd.org/en/publications/2019/03/equal-access-to-justice-for-inclusive-growth_a69ac7da.html.

Organisation for Economic Co-operation and Development. "AI Principles." (2024). https://oecd.ai/en/ai-principles.

Park, Joonsuk, and Claire Cardie. "Identifying Appropriate Support for Propositions in Online User Comments." In *Proceedings of the First Workshop on Argumentation Mining*, edited by Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, 29–38. Baltimore, MD: Association for Computational Linguistics, 2014. https://doi.org/10.3115/v1/W14-2105.

Perry, Stephen R. "Judicial Obligation, Precedent and the Common Law." *Oxford Journal of Legal Studies* 7, no 2 (1987): 215–257. https://doi.org/10.1093/ojls/7.2.215.

Prakken, Henry. *Logical Tools for Modelling Legal Argument.* Dordrecht: Springer, 1997.

Raz, Joseph. "Why Interpret?" *Ratio Juris* 9, no 4 (1996): 349–363. https://doi.org/10.1111/j.1467-9337.1996.tb00251.x.

Rescala, Paula, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. "Can Language Models Recognize Convincing Arguments?" *arXiv*, October 3, 2024. https://doi.org/10.48550/arXiv.2404.00750.

Roumeliotis, Konstantinos I, and Nikolaos D Tselikas. "ChatGPT and Open-AI Models: A Preliminary Review." *Future Internet* 15, no 6 (2023): 192. https://doi.org/10.3390/fi15060192.

Savelka, Jaromir, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. "Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise?" In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 117–123. 2023. https://doi.org/10.1145/3587102.3588792.

Sejnowski, Terrence J. *The Deep Learning Revolution*. Cambridge, MA: MIT Press, 2018.

Summers, Robert S. "Two Types of Substantive Reasons: The Core of a Theory of Common-Law Justification." *Cornell Law Review* 63, no 5 (1977): 707–788.

Thomson Reuters. "AI-Powered Legal Research: Where Legal Research Meets Generative AI." *Thomson Reuters Law Blog*, November 15, 2023. https://legal.thomsonreuters.com/blog/legal-research-meets-generative-ai.

Tu, S Sean, Amy Cyphert, and Samuel J Perl. "Artificial Intelligence: Legal Reasoning, Legal Research and Legal Writing." *Minnesota Journal of Law, Science & Technology* 25, no 2 (2024). https://scholarship.law.umn.edu/cgi/viewcontent.cgi?article=1566&context=mjlst.

Velarde, G. "Artificial Intelligence and Its Impact on the Fourth Industrial Revolution: A Review." *ArXiv*, 2020. https://doi.org/10.48550/arXiv.2011.03044.

Villasenor, John. "Generative Artificial Intelligence and the Practice of Law: Impact, Opportunities, and Risks." *Minnesota Journal of Law, Science & Technology* 25 (2024). https://scholarship.law.umn.edu/cgi/viewcontent.cgi?article=1563&context=mjlst.

Walton, Douglas. "Are Some Modus Ponens Arguments Deductively Invalid?" *Informal Logic* 22, no 1 (2002). https://doi.org/10.22329/il.v22i1.2571.

Wang, Boshi, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. "Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, 2717–2739. Toronto: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.acl-long.153.

Wang, Yuxia, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. "Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs." Preprint submitted 2023. https://arxiv.org/abs/2308.13387.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35 (2022): 24824–24837.

Wiratunga, Nirmalie, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. "CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering." *arXiv*, April 4, 2024. https://arxiv.org/abs/2404.04302v1.

World Justice Project. "Global Insights on Access to Justice 2019." https://worldjusticeproject.org/our-work/research-and-data/global-insights-access-justice-2019.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. "A Survey of Large Language Models." Preprint submitted 2023. https://arxiv.org/abs/2303.18223.

**Legal Cases**

*Carmichael Rail Network Pty Ltd v BBC Chartering Carriers GmbH &Co KG* [2024] HCA 4, 14 February 2024.
*Harvey v Minister for Primary Industry and Resources* [2024] HCA 1, 7 February 2024.
*Ismail v Minister for Immigration, Citizenship and Multicultural Affairs* [2024] HCA 2, 7 February 2024.
*Lesianawai v Minister for Immigration, Citizenship, and Multicultural Affairs* [2024] HCA 6, 6 March 2024.
*The King v Rohan* [2024] HCA 3, 14 February 2024.
*Xerri v The King* [2024] HCA 5, 6 March 2024.