GenAI as an International Lawyer:
A Case Study with the Jessup International Law Moot Court

Damien Charlotin[1] and Niccolò Ridi[2]

This paper investigates the capacity of Generative Artificial Intelligence, specifically Large Language Models, to craft compelling international legal arguments. We tested the performance of two popular models, Gemini 2.0 and GPT4o, in the Jessup International Law Moot Court Competition, generating ten complete written memorials with minimal human intervention. With the organisers' blessing, these AI-generated memorials were anonymously added to the pool of submissions and evaluated by judges, who remained unaware of their origins, providing a unique benchmark against human-produced work. Our results demonstrate that LLM-generated memorials consistently achieve average to superior scores, with some submissions receiving exceptional praise and near-perfect ratings. However, a detailed analysis of judges' qualitative feedback reveals persistent shortcomings of LLMs, notably factual inaccuracies, hallucinated citations, and superficial legal analysis. This study systematically identifies the current strengths and limitations of GenAI in legal argumentation, and critically informs best practices in prompt engineering, human-AI collaboration strategies, and emerging regulatory policies for legal education and practice.

[1] HEC Paris. damien.charlotin@gmail.com.
[2] King's College London. niccolo.ridi@kcl.ac.uk.

I. INTRODUCTION

The advent of Generative Artificial Intelligence ("GenAI") has shaken numerous professional domains. In particular, fields that rely on the articulation of text and language cannot are deeply impacted by Artificial Intelligence ("AI") that is capable of structuring information gleaned from text and producing coherent, complex language outputs.[3] Legal practice provides an exacting benchmark for these technologies, since the nature of legal writing – precise, formalized, and deeply reliant on authoritative sources – tests the limits of GenAI in terms of linguistic finesse, contextual sensitivity, and accuracy.[4]

In particular, contentious legal settings require nuanced rhetorical strategies coupled with an impeccable grasp of facts, legislation, and precedents, areas in which AI has shown both remarkable promise and notable limitations. Reflecting these developments, a recent survey of the international arbitration field, for instance, revealed that 90% of respondents expected to use AI as part of their practice.[5] Seeking to answer this demand, several legal startups promise to "Unlock the power of effortless legal argumentation",[6] or to "[m]aximise efficiency with legal AI".[7] Early indications suggest that LLMs hold promise in streamlining legal research, drafting, and argumentation, potentially democratizing access to legal services and levelling the playing field.[8] However, concerns persist regarding the accuracy and reliability of AI-generated legal arguments, and the need for human oversight to ensure ethical and responsible use. A rigorous and comprehensive investigation into these aspects is crucial to harness the benefits of LLMs in the legal field while mitigating potential risks.[9]

---

[3] D. Charlotin, "Large Language Models and the Future of Law" (August 22, 2023). Available at SSRN: https://ssrn.com/abstract=4548258 or http://dx.doi.org/10.2139/ssrn.4548258; Sai Krishnan Mohan, 'Management consulting in the Artificial Intelligence – LLM era' (2024) 7(1) *Management Consulting Journal* 9–24.

[4] Sean Tu, Amy Cyphert and Sam Perl, 'Artificial Intelligence: Legal Reasoning, Legal Research and Legal Writing' (2024) 25 *Minnesota Journal of Law, Science and Technology* 105, WVU College of Law Research Paper No 2024-012 https://ssrn.com/abstract=4817765, p. 124.

[5] White & Case, '2025 International Arbitration Survey – The path forward: Realities and opportunities in arbitration' (April 2025, available here).

[6] See AI Legal Argument Generator, Taskade, https://www.taskade.com/generate/legal/legal-argument (last visited May 26, 2025).

[7] See Uncover, https://www.uncoverlegal.com/ (last visited May 26, 2025)..

[8] *See* Erik Brynjolfsson, Danielle Li, & Lindsey Raymond, *Generative AI at Work*, 140 Q.J. ECON. 889, 889–942 (2025), https://doi.org/10.1093/qje/qjae044 (describing more beneficial effects from AI use for novices and lower-skilled workers).

[9] Anticipating such research, Runar Hilleren Lie & Malcolm Langford, *The Computational Turn in International Law*, 93 NORDIC J. INT'L L. 38, 61 (2024), pointed out that "The experimentation with the potential of leveraging llms for greater methodological power and sophistication is only now beginning."

In this paper, we conduct an experiment to shed some light on the potential and pitfalls of using GenAI in law, and in international law in particular. With the assistance of widely-used LLMs, and the blessing of the organisers, we generated a set of 10 written submissions for the Jessup International Law Moot Court Competition – a competition celebrated for its rigour and connection with the practice of international law. These memorials were entered into the pool and graded by judges unaware of their source or drafting process. While all memorials were generated in light of the Jessup Problem and the Basic Materials provided by the Jessup organisation, some benefitted from additional context reflecting the result of deeper legal research. But while we lightly polished the formatting[10] of the generated texts, we did nothing to alter their substance.

As recounted below, this batch of AI-generated memorials largely scored average grades; some memorials, however, scored highly enough to appear in the top 5% of all submissions. Example comments from the judges included the sentiment that "From the batch of memorials that were assigned to me, I rated yours the highest" (grade of 97/100), or "Excellent submission! The issues in dispute were precisely identified, and the arguments well researched and well presented" (grade of 96/100). Other judges, however, identified markers (such as factual inconsistencies, citation issues, and repetitions) that highlight the limitations of LLMs in this exercise.

By leveraging the Jessup's detailed scoring system, which provides granular scores for elements such as knowledge of the law, legal analysis, clarity, and adherence to formal requirements, we are also able to assess how AI performs in direct comparison with human teams. The results – in terms of the grades and their various components, but also a *corpus* of comments from the judges – allow us to identify where AI-generated arguments shine, and where they remain lacking.

This experiment is significant for several reasons. First, it provides the first systematic, blind evaluation of frontier commercial models' ability to generate end-to-end complex legal submissions, moving beyond the focus on legal exams or component tasks in the existing empirical literature. As such, it provides a systematic evaluation of AI's current abilities in a context which values characteristics such as precision, creativity, and compliance with rules. Second, it sheds light on the challenges and opportunities associated with using AI in legal argumentation, identifying specific areas where LLMs excel and others where they struggle. Third, by experimenting with different levels of human-AI collaboration, this study contributes to understanding how best to harness AI's potential in professional legal contexts.

Our findings have broader implications. They provide insights into how GenAI can assist international lawyers and legal practitioners in drafting arguments, performing research, and improving the overall efficiency of legal work. Furthermore, they contribute to the

[10] For instance, we removed tell-tale signs of AI-generated texts, such as em dashes, or of copying and paste, such as straight quote marks.

development of more effective prompting strategies and fine-tuning methods for AI tools tailored to legal applications. Below, we introduce the role LLMs can play in generating texts and arguments for contentious legal proceedings (**I**). We then describe the design of our study (**II**), before discussing our findings (**III**).

A. Large Language Models and the Future of the Law

1. Introduction to Large Language Models in the Legal Context

Large Language Models (LLMs), such as OpenAI's GPT series and Google's Gemini, represent a transformative leap in artificial intelligence. The vast neural networks that power them were trained on diverse textual datasets, enabling LLMs to generate coherent and contextually appropriate text across a wide array of applications. In the legal context, LLMs offer capabilities ranging from drafting contracts and summarizing case law to generating persuasive arguments and even simulating courtroom dialogue.[11]

The journey to LLMs' current sophistication reflects the broader evolution of AI in legal practice.[12] Early applications, such as rule-based expert systems between the 1970s and 1990s, were limited to answering narrowly defined questions or performing predefined tasks.[13] These systems required exhaustive programming of specific legal rules and were inflexible in adapting to novel scenarios. The advent of machine learning in the 2000s, particularly natural language processing ("NLP"), enabled more dynamic tools, such as keyword-based legal research platforms and e-discovery software. However, these tools often relied on basic string search or shallow textual patterns, and entirely lacked the ability to generate original legal arguments.

The advent of LLMs thus mark a paradigm shift by combining the strengths of deep learning with vast contextual understanding (or, at least, the appearance of such understanding). For example, GPT-4 can draft a detailed brief on tort law or analyze ambiguities in international treaties by leveraging its pretraining on diverse legal as well as non-legal *corpora*. Unlike earlier systems, modern LLMs require minimal explicit programming to engage with

---

[11] Daniel Schwarcz et al., *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice* (Minnesota Legal Studies Research Paper No. 25-16, 2 March 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111 (last visited May 26, 2025).

[12] *See generally* Trevor Bench-Capon et al., *A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law*, 20 ARTIFICIAL INTELLIGENCE & L. 215 (2012), https://doi.org/10.1007/s10506-012-9131-x (last visited May 26, 2025).

[13] *See, e.g.*, L. T. MCCARTY, REFLECTIONS ON "TAXMAN": AN EXPERIMENT IN ARTIFICIAL INTELLIGENCE AND LEGAL REASONING, 90 HARV. L. REV. 837, 837–893 (1977); *see generally* R. E. SUSSKIND, EXPERT SYSTEMS IN LAW: JURISPRUDENTIAL INQUIRY (1987).

complex, nuanced questions. This ability to process and synthesize information at scale positions LLMs as uniquely suited to augment legal analysis and drafting.

Despite their potential, LLMs face inherent limitations. Their outputs depend heavily on the quality of prompts and the breadth of their training data. Moreover, LLMs arguably lack true "comprehension"[14] - fundamentally, their responses are based on statistical predictions based on patterns in the data they were trained on. As a result, they can sometimes produce results that appear plausible, but are in fact far from correct, a phenomenon often referred to as "hallucination."[15]

In the legal domain, where precision and accuracy are paramount, such errors can have significant consequences. Notorious examples from legal practice include the citation of non-existent legal authorities,[16] a failure which is revealing of two interrelated limitations of LLMs. First, while LLMs are exposed to an enormous amount of knowledge in their training, they are neither omniscient nor designed to be, and they may not have access to all the relevant information.[17] Second, because they are trained to produce results that fit specific patterns, they may in any event attempt to produce a result consistent with the typical way of justifying and grounding an assertion – in the legal domain, likely through a citation of authority, genuine or otherwise.

2. Capabilities of LLMs in Legal Applications

Recognizing their burgeoning potential, researchers have begun to rigorously assess LLM performance in the legal domain through diverse methodologies and in various contexts.[18]

---

[14] This term, like "consciousness", should of course be caveated to the extent that its definition is in fact unclear.

[15] *See generally* MATTHEW DAHL, VARUN MAGESH, MIRAC SUZGUN, & DANIEL E. HO, LARGE LEGAL FICTIONS: PROFILING LEGAL HALLUCINATIONS IN LARGE LANGUAGE MODELS, 16 J. LEGAL ANALYSIS 64, 66 (2024).

[16] DANIEL SCHWARCZ et al., *supra* note [**crossref**], at 13. For an empirical tally of this issue in legal proceedings, see Damien Charlotin, "AI Hallucinations Cases", database available at: https://www.damiencharlotin.com/hallucinations/.

[17] See *ZZaman v. HMRC*, [2025] UKFTT 00539 (TC), 3 April 2025, para. 29 ("This highlights the dangers of reliance on AI tools without human checks to confirm that assertions the tool is generating are accurate. Litigants using AI tools for legal research would be well advised to check carefully what it produces and any authorities that are referenced. These tools may not have access to the authorities required to produce an accurate answer, may not fully "understand" what is being asked or may miss relevant materials. When this happens, AI tools may produce an answer that seems plausible, but which is not accurate.")

[18] At the outset, it bears emphasizing that, by contrast with other fields - such as maths and coding for instance - testing LLMs and the legal field is inherently harder: the research community lacks comprehensive benchmarks that allow a thorough testing of LLMs answers to legal conundra.

One initial baseline for assessing LLM legal capabilities has been their performance on law school exams and bar exams. Early in 2023, Choi and others tested the original ChatGPT (based on GPT-3.5) on four University of Minnesota law school final exams. ChatGPT's answers (95 multiple-choice and 12 essay questions) received an average grade of C+, a low but passing performance.[19] In other words, GPT-3.5 performed comparably to as a bottom-tier law student – competent on basic issues but often failing to capture crucial points of analysis.

However, the field witnessed a significant inflection point with the advent of GPT-4. OpenAI's internal tests showed GPT-4 scoring near the top 10% on a simulated bar exam, a stark contrast to GPT-3.5's bottom 10% performance. [20] While independent analyses later adjusted GPT-4's percentile to potentially the 60th-70th range against actual test-taker distributions, the trajectory was undeniable: LLMs were rapidly improving in their ability to address formal legal knowledge and reasoning tasks.[21]

Beyond standardized tests, researchers have examined LLM assistance in law school settings. Choi and Schwarcz conducted a controlled experiment where J.D. students took exams with and without GPT-4's help. While GPT-4 significantly enhanced performance on simpler multiple-choice questions, it offered minimal advantage for complex essays demanding sophisticated legal analysis. Moreover, the benefits were inversely correlated with student aptitude; lower-performing students experienced substantial gains, whereas top

---

This stems in no small part from the very indeterminacy of the law, as well as its requirements in terms not only of legal output, but also of a legal *reasoning* to get there. But see, e.g., Guha et al. 'LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models', available here.

[19] Jonathan H. Choi, Kristin E. Hickman, Amy Monahan & Daniel Schwarcz, *ChatGPT Goes to Law School*, 71 J. LEGAL EDUC. 387 (2022).

[20] OPENAI, GPT-4 (Mar. 14, 2023), https://openai.com/index/gpt-4-research/ (last visited May 26, 2025).

[21] Other experiments have been carried out the on the ability of LLMs to solve legal tasks related to bar exam taking. *See, inter alia*, MICHAEL JAMES BOMMARITO & DANIEL MARTIN KATZ, *Gpt Takes the Bar Exam*, SSRN ELECTRONIC JOURNAL (Dec. 2022), https://doi.org/10.2139/ssrn.4314839 (last visited May 26, 2025); DANIEL MARTIN KATZ, MICHAEL JAMES BOMMARITO, SHANG GAO & PABLO ARREDONDO, *Gpt-4 Passes the Bar Exam*, 382 PHIL. TRANSACTIONS ROYAL SOC'Y A (2024), https://doi.org/10.2139/ssrn.4389233 (last visited May 26, 2025); *see also*, on fine-tuning a model for this very purpose, REAN FERNANDES, ANDRÉ BIEDENKAPP, FRANK HUTTER & NOOR AWAD, *A Llama Walks into the 'Bar': Efficient Supervised Fine-Tuning for Legal Reasoning in the Multi-state Bar Exam*, https://doi.org/10.48550/arXiv.2504.04945 (last visited May 26, 2025).

students saw a slight decline, suggesting AI might act as an educational leveller but potentially blunt the edge of high-achievers.[22]

The same study also assessed GPT-4's autonomous exam performance. With rudimentary prompting, its output was mediocre (C-grade equivalent). However, with optimized prompting, GPT-4 surpassed not only the class average but also the average AI-assisted student. This indicated that, for certain structured tasks, a well-guided LLM might not just assist, but potentially obviate the need for direct human involvement, hinting at a future where "humans are removed from the loop" for specific applications.[23]

Further studies reinforced this emerging profile of LLM capabilities. Armin Alimardani, for example, found that while GPT-4 underperformed students on Australian law exam questions requiring intricate legal reasoning, it excelled in open-ended essays emphasizing composition and general knowledge. This pattern – proficiency in articulation and straightforward issue handling, contrasted with limitations in deep, domain-specific analysis – became a recurring theme in early LLM evaluations.

The focus then expanded towards practical legal tasks, such as drafting. Trozze et al. (2024) conducted a notable experiment where mock jurors evaluated legal complaints, some drafted by human lawyers and others by GPT-3.5 (an earlier, less capable model).[24] Remarkably, jurors showed no significant difference in decisions, confidence, or their assessment of whether legal claims were proven, regardless of whether the pleading was human- or AI-authored. This suggested that even less advanced LLMs could produce litigation documents persuasive to laypersons, potentially rivalling human lawyers in specific drafting contexts. This finding also resonates with observations of a "more human than human" effect, where LLM-generated text can sometimes be indistinguishable from, or even preferred over, human writing. [25]

Recent experimental evidence supports the view that large language models exhibit increasingly competent performance on legal tasks when equipped with either enhanced

---

[22] Jonathan H. Choi & Daniel Schwarcz, *AI Assistance in Legal Analysis: An Empirical Study*, 73 J. LEGAL EDUC. (forthcoming 2024). This research should be put in parallel with subsequent experiments in non-legal fields that found a discrepancy in how AI improves performance, suggesting that the nature of the task matters tremendously: in particular, in complex, knowledge-intensive domains, AI appears to be an amplifier of existing expertise rather than a universal leveller. See NICHOLAS OTIS et al., *The Uneven Impact of Generative AI on Entrepreneurial Performance* (Feb. 27, 2024), https://ssrn.com/abstract=4671369 (last visited May 26, 2025).

[23] Ibid.

[24] Arianna Trozze, Toby Davies & Bennett Kleinberg, *Large Language Models in Cryptocurrency Securities Cases: Can a GPT Model Meaningfully Assist Lawyers?*, ARTIFICIAL INTELLIGENCE AND LAW, 1–47 (2024).

[25] See, Cameron R. Jones & Benjamin K. Bergen, *Large Language Models Pass the Turing Test*, arXiv preprint arXiv:2503.23674 (2025).

reasoning capacity or domain-specific grounding. In particular, Schwarcz et al. (2025) demonstrate that both OpenAI's *o1-preview* (a reasoning-augmented LLM) and *Vincent AI* (an LLM-based tool augmented with the function to retrieve relevant information for the task at hand) significantly improve the quality and efficiency of legal outputs produced by upper-level law students.

Concurrently, the adaptability of LLM to the legal domain is quickly becoming a selling point: in releasing the GPT-4.1 model, OpenAI mentioned specifically how "[they] trained GPT-4.1 to reliably attend to information across the full 1 million context length[26] [... and]to be far more reliable than GPT-4o at noticing relevant text, and ignoring distractors across long and short context lengths" as "[l]ong-context understanding is a critical capability for applications across legal, coding, customer support, and many other domains" and specifically cited Thomson Reuters' experiments with their CoCounsel platform as showing the prowess of their newly-released model.[27]

These technological advancements and empirical findings have not gone unnoticed by the legal profession. A late 2023 survey by Bryan Cave Leighton Paisner (BCLP) of arbitral dispute practitioners revealed a profession actively engaging with AI.[28] While 90% reported awareness, current adoption centered on ancillary tasks like document translation, review, formatting, and summarization – functions well-suited to arbitration's document-intensive nature. This indicates a pragmatic, yet cautious, approach to integrating AI, particularly for high-stakes adjudicative functions, reflecting an ongoing assessment of AI's evolving capabilities and risks.

3. Challenges and Limitations of LLMs in Law

Despite the recognized utility of LLMs for ancillary tasks, significant skepticism persists regarding their application in core legal functions like generating arguments or decisions. Just 24% of respondents in the BCLP survey had used AI for drafting documents for arbitration, and a majority (53%) believed AI tools should not generate text for legal arguments or submissions. A striking 88% of respondents expressed concern over AI

[26] "Context length" refers to the amount of text (expressed not in words, but in word parts known as tokens) that can be included in the prompt of LLM (and thus, participate to the prediction of the generated output).

[27] OpenAI, Introducing GPT-4.1 in the API, 14 April 2025, https://openai.com/index/gpt-4-1/.

[28] Specifically, 90% of respondents were aware of AI tools for legal tasks, and many had already used them in ancillary tasks: the most common current uses were translation of documents (37% of respondents), document review and e-discovery (30%), formatting and editing texts (30%), and summarizing documents (24%). *See* BCLP ARBITRATION SURVEY 2023 AI IN IA: THE RISE OF MACHINE LEARNING, BCLP, https://www.bclplaw.com/en-US/events-insights-news/bclp-arbitration-survey-2023.html#:~:text=The%20topic%20for%20our%202023,of%20AI%20in%20international%20arbitration. (last visited May 26, 2025).

.

"hallucinations" – the fabrication of fictitious facts or legal precedents – highlighting a critical barrier to trust.

This caution highlights crucial limitations of LLMs in legal contexts. Foremost among these is the challenge of factual accuracy and the propensity for hallucination. Although LLMs can produce coherent and persuasive narratives, their tendency to generate spurious information represents a severe risk in a domain where precision is paramount. Indeed, an increasing number of legal disputes have been compromised by AI-generated inaccurate information.[29] The high premium on accuracy in legal proceedings means that any error can have serious ramifications, potentially undermining case integrity or leading to erroneous judicial decisions. In essence, the legal domain exhibits an asymmetry of risk: the cost of a false positive (a hallucination) far outweighs that of a false negative (omitting a relevant point).

Contextual understanding remains another critical concern. While LLMs generate text that superficially appears legally sound, they often struggle with the inherent subtleties of legal language and complex doctrinal nuances.[30] Jurisdiction-specific laws and regulations require an appreciation for differences that can hinge on minor semantic or contextual shifts; however, current models tend to generalize across entire domains, let alone legal systems, potentially overlooking these fine distinctions. This lack of precise contextual sensitivity risks misinterpretation and potential misapplication of legal principles. This deficit contributes to what Tu, Cypher, and Perl (2024) term an "alignment gap between the performance of certain AI systems and the needs of lawyers."[31]

Bias and fairness also stand out as significant challenges. LLMs, having been trained on extensive datasets that mirror the biases embedded in legal literature and broader societal discourse, can inadvertently perpetuate discriminatory notions.[32] Their outputs might contain subtle bias or reinforce existing stereotypes,[33] leading to outcomes that are not only

---

[29] *See generally* Damien Charlotin, *AI Hallucination Cases*, https://www.damiencharlotin.com/hallucinations/ (last visited May 26, 2025).

[30] *See* SALVATORE GIORGI et al., *Explicit and Implicit Large Language Model Personas Generate Opinions but Fail to Replicate Deeper Perceptions and Biases*, ARXIV E-PRINTS (2024): arXiv-2406. (Stating that "LLMs lack the intrinsic cognitive mechanisms of human thought, while capturing the statistical patterns of how people speak, which may restrict their effectiveness in complex social science applications.")

[31] Shine (Sean) Tu, Amy Cyphert & Sam Perl, *Artificial Intelligence: Legal Reasoning, Legal Research and Legal Writing*, 25 MINN. J.L. SCI. & TECH. 105, 105–25 (2024).

[32] See Lin, Luyang, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. "Investigating bias in llm-based bias detection: Disparities between llms and human perception." arXiv preprint arXiv:2403.14896 (2024).

[33] See for example Wan, Yixin, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters." arXiv preprint arXiv:2310.09219 (2023).

ethically problematic but also legally detrimental, particularly in adversarial settings where impartiality is essential.

All this without even mentioning the ethical and professional implications of deploying LLMs in legal practice, which are manifold. Issues of confidentiality, professional responsibility, transparency and disclosure, accountability and liability, are all bound to be up for debate – but beyond the scope of this paper.[34]

Interestingly, initial misgivings regarding the creative and original capacities of LLMs in legal argumentation appear to be evolving. While early critiques often dismissed LLMs as mere "stochastic parrots" incapable of genuine inventive breakthroughs,[35] the performance of more recent models has begun to challenge this perception. For instance, studies assessing creativity in fields like poetry or meme generations[36] suggest that audiences often rate unlabelled AI-generated work as more pleasing or innovative than average human-produced content. While not directly analogous to legal reasoning, these findings hint at an evolving potential for AI to contribute novel perspectives.[37]

Nevertheless, the caution voiced by Tu, Cyphert, and Perl regarding the potential impact on human creativity remains pertinent: they argue that "[w]hen lawyers skip the initial drafting phase of their writing, instead choosing to immediately use a generative AI tool, they potentially undercut their own creativity and could end up with a written product that makes weaker legal arguments than they would have developed by doing the hard work of thrashing

---

[34] Google's privacy policy for Gemini, for example, contains the following stern warning: "Please don't enter confidential information in your conversations, or any data you wouldn't want a reviewer to see or Google to use to improve our products, services and machine learning technologies." This warning highlights a significant privacy concern with LLMs. When users input confidential information into these systems, even if they believe their conversation is private, that data can be exposed to human reviewers working to improve the system. These reviewers may include third-party contractors who access conversations to annotate and evaluate them, albeit with certain privacy measures in place. Additionally, the confidential information could potentially be incorporated into training data for future model iterations, meaning sensitive details might inadvertently influence responses to other users or be used in ways the original user never intended. This creates a risk – minor, perhaps, at the level of probability, but hardly tolerable – that proprietary, personal, or otherwise sensitive information could be exposed beyond its intended audience. *See* GEMINI APPS PRIVACY HUB, GOOGLE SUPPORT, https://support.google.com/gemini/answer/13594961?hl=en-GB (last updated May 23, 2025, last visited May 26, 2025).

[35] See Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *in* PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610 (ACM 2021), https://doi.org/10.1145/3442188.3445922.

[37] *See* Behnam Mohammadi, *Creativity Has Left the Chat: The Price of Debiasing Language Models*, arXiv (June 8, 2024), https://arxiv.org/abs/2406.05587 (last visited May 26, 2025) (highlighting that the difference between mere mimicry and genuine creative synthesis can hinge on technical factors such as the structure of the training data and the sophistication of the prompt engineering).

around in an initial draft."[38] The implications of these technological developments for legal education, professional training, and the very nature of legal craftsmanship are indeed likely to be at the forefront of the profession's debates in the coming years.[39]

4. Implications for Legal Practice

The foregoing considerations suggest that the integration of LLMs into legal practice necessitates a fundamental re-evaluation of what constitutes legal expertise. As advanced models increasingly handle routine drafting, legal research, and initial argument generation, the lawyer's function is poised to evolve from being the primary executor of technical tasks to that of a strategic and ethical overseer[40] One can readily envision future lawyers using LLMs to produce first drafts of complaints, motions, or briefs, subsequently editing and refining the content – leveraging AI's strength in language generation while applying human judgment to ensure legal accuracy, strategic alignment, and ethical compliance.

Consequently, mastery in prompt engineering and the critical validation of AI outputs may become as indispensable as traditional legal knowledge. This evolving role may demand not only familiarity with doctrinal principles but also an acute understanding of AI's operational parameters, transforming legal expertise into a hybrid discipline of law and technology.

The promise of improved access to justice also emerges as a transformative theoretical implication. LLMs can lower entry barriers by significantly reducing the cost and time required to perform essential legal tasks, thereby democratizing access to legal services for individuals and small firms. This dynamic fosters a legal landscape where quality legal support is not solely the preserve of large law firms, potentially narrowing the gap between disparate segments of society.

Despite these potential advantages, human oversight remains non-negotiable. The research underlines that even when AI-generated outputs appear persuasive and well-crafted, the final adjudication of legal arguments must reside with trained professionals who can detect subtle inaccuracies or misinterpretations. Practitioners must therefore maintain an active supervisory role, ensuring that technology serves as an aid rather than a substitute for the principled discretion of human judgment.

Furthermore, the significance of training – both for the AI models and the humans utilizing them – along with sophisticated prompt engineering, cannot be overstated. The quality of AI

---

[38] Sean Tu, Amy Cyphert & Sam Perl, *Artificial Intelligence: Legal Reasoning, Legal Research and Legal Writing*, 25 MINN. J.L. SCI. & TECH. 105, 124 (2024).

[39] Amanda Head & Sonya Willis, *Assessing Law Students In A GenAI World To Create Knowledgeable Future Lawyers*, 31 INT'L J. LEGAL PROF. 293 (2024). *See also* Sarah O'Connor, *A white-collar world without juniors?*, FIN. TIMES, Mar. 25, 2025.

[40] This development had been anticipated: *see, e.g.*, KATIA FACH GÓMEZ, THE TECHNOLOGICAL COMPETENCE OF ARBITRATORS (Springer 2023).

outputs is heavily contingent on the inputs provided by users. Developing robust strategies for prompt construction is therefore key to optimizing LLM performance in legal contexts.[41] Thus, mastering prompt engineering enhances the relevance and accuracy of AI-generated text, but also acts as a form of continuous learning – enabling legal practitioners to extract more valuable, context-appropriate responses from LLMs.

Our study offers empirical support for these evolving dynamics. For instance, effective prompt engineering was essential in guiding the model's responses towards legally relevant and structured arguments. Nevertheless, hallucinations did occur, especially when generating responses that required specialized knowledge or nuanced common-sense reasoning not explicitly present in the provided context materials. Iterative refinement of prompts and the strategic incorporation of additional, targeted information (simulating user feedback and domain expertise) were crucial in improving the overall quality and accuracy of the generated memorials, underscoring the symbiotic relationship required between human guidance and AI capability

B. Design of the Study

The design of this study leverages the Jessup International Law Moot Court Competition as a rigorous testing environment, in which we can test how GenAI fare with a level of granular detail (given the competition's detailed grading guidelines). We first detail the competition's framework, scoring mechanisms, and relevance for assessing AI capabilities in legal drafting – especially in view of the AI policies recently deployed by moot competitions (*1*). Following this contextual background, we describe the specific process of generating the experimental memorials using LLMs (*2*).

1. Moot Court Competition

a) The Philip C. Jessup International Law Moot Court Competition

The Jessup is, according to their own data, the world's largest moot court competition, and a prominent venue amongst the "grand slam" of international moots, characterised by their leading character "in their respective fields, marked by the scale, highest level of prestige,

[41] As put by the UK court in ZZaman v. HMRC, [2025] UKFTT 00539 (TC), 3 April 2025: "There is no reliable way to stop [hallucinations from AI tools], but the dangers can be reduced by the use of clear prompts, asking the tool to cite specific paragraphs of authorities (so that it is easy to check if the paragraphs support the argument advanced), checking to see the tool has access to live internet data, asking the tool not to provide an answer if it is not sure and asking the tool for information on the shortcomings of the case being advanced."

rich history, excellent organisation, quality and diversity of judging, and quality and diversity of participants."[42]

Each year, the Jessup brings together law students from over 700 universities across more than 100 countries to engage in simulated proceedings before the International Court of Justice (ICJ).[43] At the heart of the competition lies the 'Compromis' – a hypothetical case crafted by leading scholars of international law that serves as the springboard for the entire competition. The Compromis, the writing of which is seen by many as an honour in itself, presents a fictional dispute between two States, carefully constructed to address complex, if classic, issues of public international law.

The Compromis typically hinges on four main, interconnected issues for which there is no single "correct" solution in light of the current state of development of international law. Thus, the issues offer arguments for both sides of the debate – that is to say, Applicant and Respondent. In its 2025 edition, the Jessup Compromis involved the fictional nations of Ambrosia and Rovinia, and included the following issues:

- Jurisdiction *ratione temporis* over acts and conducts that took place before the relevant treaty entered into force;
- Jurisdictional authority over an official in relation to crimes committed abroad;
- Legality of maritime baselines frozen in time, in areas affected by climate change; and
- Legality of the seizure of state property abroad.

The competition unfolds in two distinct phases: the written and the oral, with the former phase being particularly crucial, since it forms the foundation upon which oral arguments are later built. Each team is required to produce two memorials for Applicant and Respondent respectively. These memorials must cover all four issues presented in the Compromis within a strict word limit of 9,500 words, and their drafting process is rigorous and time-consuming, often spanning several months between the Compromis's release, typically in early September, and the written phase deadline, in mid-January of the following year. Students must engage in extensive research, critically analyse complex legal issues, and construct persuasive arguments for both sides of the dispute. This dual perspective, in principle, encourages students' ability to anticipate and counter opposing arguments.

The strict word limit imposed on memorials – 9,500 words – presents a unique challenge: students must learn to articulate complex legal arguments concisely and effectively, a skill

---

42 *See* https://internationalmoots.org/grand-slams (last visited May 26, 2025).

43 See *See* DAVID M. SCOTT & ULLA SOIRILA, *The Politics of the Moot Court*, 32 EUR. J. INT'L L. 1079 (2021). The 2025 edition was the largest on record, with over 900 teams participating, and 160+ proceeding to the Finale in Washington, D.C.

that is highly valued in legal practice. This constraint also encourages teams to prioritise their strongest arguments and present them in the most compelling manner possible.

One of the most distinctive features of the Jessup is its intended commitment to balance. As the competition organisers note, "judges want to focus on your team's ability to prepare and present good legal arguments, rather than decide which team has the 'winning' legal argument". To this end, the Compromis is crafted with intentional ambiguities, allowing both parties to use particular facts to argue their case effectively. This focus on technical legal argumentation has been critiqued by some scholars, who argue that the competition structure can render "the unequal power and negotiating positions of Global South states invisible through the fiction of equal legal opponents".[44]

b) Scoring

With its well-defined rules and scoring system, the Jessup provides a unique setting for evaluating the performance of AI-generated memorials. The competition's rules on scoring and penalties offer a rare opportunity to compare legal writing, research, and argumentation against a quantifiable framework for assessing their quality, allowing us to directly compare the performance of our AI-generated memorials with those drafted by human participants.

Specifically, the Jessup's scoring system, as detailed in Rule 10, provides a multi-dimensional measure of memorial quality. Each memorial is evaluated by at least three judges, on a 100-point scale that mandates consideration of a memorial's knowledge of the law, legal analysis, research, clarity, organization, and writing style. Thus, by comparing the scores received by our AI memorials with those of human-drafted memorials, we can therefore gain a clear understanding of the strengths and weaknesses of LLMs in producing high-quality legal writing.

Furthermore, the Jessup's rules provide for specific violations and penalties. The rules specify a range of penalties for various infractions, including exceeding word limits, improper citation, and plagiarism (Rule 11.1). By examining the types and frequency of penalties incurred by our AI-generated memorials, we can identify specific areas where LLMs struggle to meet the competition's standards. For instance, penalties for improper citation (Rule 5.13) could highlight challenges in accurately attributing sources and adhering to specific citation formats.

The detailed scoring system and the comprehensive rules on violations allow for a granular assessment of the quality of our AI-generated memorials. By analysing the scores and penalties received, we can pinpoint the specific areas where LLMs excel and where they fall

[44] Ibid.

short, providing valuable insights into the capabilities and limitations of AI in legal writing and argumentation.

c) Regulation of AI in Moots

Our hypothesis that moot competitions are fertile ground to test the capabilities of LLMs finds anecdotal support in the increasing concern of competition organisers for the use of AI. Needless to say, the power of LLMs to perform as well as lawyers in some contexts, or at least to convincingly imitate the conventions of legal writing, raises significant questions for competitions that include a written component, such as the Jessup.

This is so, *inter alia*, because the very nature of AI-generated text makes it difficult to distinguish from human-written text. Unlike traditional plagiarism, which involves copying existing text, AI models generate new text that is often grammatically correct and semantically coherent. Jakesh et al. demonstrated that humans cannot reliably detect AI-generated content, and are actually biased in a way that facilitates AI sounding "more human than human", as noted above.[45] This makes it challenging to develop reliable detection methods based on textual features alone.

As Sadasivan and others note, "identifying AI-generated text is a critical problem to avoid its misuse [...] However, deploying vulnerable detectors may not be the right solution to tackle this issue since it can cause its own damages, such as falsely accusing a human of plagiarism."[46] Furthermore, AI content detection tools often exhibit bias against non-native English speakers. In one study, essays in English written by Chinese students were branded by text-analysis tools as being generated by artificial intelligence 61% of the time.[47] This bias stems from the fact that these tools are typically trained on large datasets of English text written by native speakers. As a result, they may struggle to accurately assess the writing style of individuals whose first language is not English. This bias raises serious concerns about the fairness and equity of using AI content detection tools in academic settings, particularly in contexts where participants come from diverse linguistic backgrounds.

The further implications for the Jessup, and mooting in general, are profound. Students might use AI to generate initial drafts of their memorials. On the one hand, this can potentially level the playing field between teams with varying research and language capabilities, in line with studies (highlighted above) that have found that AI can act as a

---

[45] Maurice Jakesch, Jeffrey T. Hancock & Mor Naaman, *Human Heuristics for AI-Generated Language Are Flawed*, 120 PROC. NAT'L ACAD. SCI. e2208839120 (2023), https://doi.org/10.1073/pnas.2208839120.

[46] V.S. Sadasivan et al., *Can AI-Generated Text be Reliably Detected?*, ARXIV, https://arxiv.org/abs/2303.11156 (last visited May 26, 2025).

[47] Christian Stokel-Walker, *Tools to Spot AI Essays Show Bias Against Non-Native English Speakers*, NEW SCIENTIST, Apr. 23, 2023, https://www.newscientist.com/article/2371004-tools-to-spot-ai-essays-show-bias-against-non-native-english-speakers/ (last visited May 26, 2025).

leveller. However, this also raises concerns about the authenticity of student work and the development of crucial legal skills. Competition organizers, including those of the Jessup, are actively considering how to adapt their rules and judging criteria to account for the potential use of AI, a topic of ongoing discussion and development within the mooting community.

Different moot court competitions are exploring various approaches to the use of AI. For example, the Vis Moot, another global competition focused on commercial arbitration proceedings, has taken a proactive stance on the use of AI. For its 31st edition, the Vis Moot has introduced new rules that aim to regulate, rather than prohibit, the use of AI tools. According to Rule 64, the Vis Moot allows the use of AI in specific contexts, including: AI-enhanced research tools for finding relevant sources; AI-enhanced proofreading tools; AI-enhanced translation tools for legal sources; AI-enhanced translation tools for text within submitted memoranda (though discouraged); AI for generating overviews or briefings on relevant topics for the team's understanding. However, Rule 65 explicitly prohibits submitting AI-generated text in memoranda (except for translations) and using the Vis Moot Problem to train AI tools. As an enforcement measure, Rule 66 requires teams to submit an "Academic Integrity and Artificial Intelligence Disclosure Statement" with their memoranda, signed by all team members and coaches.

According to some commentators,[48] this approach represents a careful balance between acknowledging AI's potential in legal practice and maintaining the academic integrity of the competition. The authors argue that by allowing AI for research purposes but prohibiting AI-generated submissions, the Vis Moot aims to ensure that students continue to develop critical legal skills while familiarizing themselves with AI tools they may encounter in their future careers. They further note that the Vis Moot's rules reflect current industry sentiment, citing a survey by Bryan Cave Leighton Paisner LLP showing 63% of arbitration practitioners support regulating, rather than banning, AI use in international arbitration. However, they did single out the challenges in regulating AI use, including the difficulty in their enforcement.

These budding approaches to AI regulation, especially in the context of legal education and competitions, would benefit from a deeper understanding of AI's capabilities. This includes a more nuanced grasp of how AI tools can be leveraged to generate legal documents, such as memorials – as illustrated in this experiment.

2. Memorials Generation

[48] Kevin Cheung & Maite Aguirre Quiñonero, *The Vis Moot's New AI Rules: Reflecting Current Sentiment & Foreshadowing Issues in Practice*, KLUWER ARB. BLOG (Dec. 12, 2023), https://arbitrationblog.kluwerarbitration.com/.

With the help of two distinct LLMs, we generated in full 10 Memorials designed to compete in the Jessup, using the Compromis as the primary prompt, along with additional instructions tailored to each experimental condition. We duplicated these memorials to gather more grades, for a total of 20 memorials.

While the results are interesting (and reviewed *infra*), the process through which we obtained these memorials yielded particular insights as well. This is why this section describes in detail our approach, with further information found in the Annex to this paper. A link to the generated memorials [shall be] included in the Annex.

a) State of the Art

In line with the current state of the art, we have followed different prompting strategies. This is so because, while advancements in LLMs have enabled these models to exhibit increasingly sophisticated reasoning capabilities across a wide range of domains,[49] their performance can vary significantly depending on the prompting strategy employed to elicit the desired behavior.[50] These strategies are worth reviewing briefly.

The first approach is known as *zero-shot prompting*, where the model is provided with instructions or a task description without any accompanying examples.[51] This method directly probes the model's inherent understanding and ability to perform the specified task based solely on the natural language prompt. In contrast, *few-shot prompting* involves the inclusion of a small set of examples (typically 1-10) that demonstrate the desired input/output format or task structure.[52] By learning to imitate these examples, the model can generalize to new instances of the task. Empirical studies have shown that performance often scales with the number of examples provided.[53]

Retrieval-Augmented Generation (RAG) has emerged as a key technique to make LLMs reliable research assistants. In RAG systems, the model is connected to a vector database of

---

[49] BROWN et al., *Language Models Are Few-Shot Learners*, 33 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 1877, 1877–1901 (2020); CHOWDHERY et al., *PaLM: Scaling Language Modeling with Pathways*, ARXIV PREPRINT ARXIV:2204.02311 (2022).

[50] P. Liu et al., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, arXiv preprint arXiv:2107.13586 (2021).

[51] BROWN et al., LANGUAGE MODELS ARE FEW-SHOT LEARNERS, 33 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 1877, 1877–1901 (2020).

[52] BROWN et al., LANGUAGE MODELS ARE FEW-SHOT LEARNERS, 33 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 1877, 1877–1901 (2020).

[53] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work?. arXiv preprint arXiv:2202.12837.

accepted sources (including legal sources) and must ground its answers in those sources, which are provided via prompts. This approach is employed by commercial tools like Lexis+ AI and Westlaw Precision AI (offered by the major legal research companies LexisNexis and Thomson Reuters). A 2024 study by Stanford researchers evaluated these tools against GPT-4 on a set of over 200 legal research questions, and their results showed that augmenting GPT-4 with a curated legal database improved accuracy, but did not eliminate errors; in particular, even the top systems still produced hallucinations in 17-33% of their answers.[54] As such, no current AI tool is truly "100% hallucination-free" despite marketing claims. More recently, evidence from randomized controlled trials has clarified the *practical impact* of RAG systems when used by human lawyers. More recently, further research established that a RAG system improved the clarity and organization of legal work across multiple tasks but showed no consistent effect on the legal accuracy of answers, and in some cases, may have even degraded it.[55]

Another approach relies on Chain-of-thought ("CoT") prompting as a particularly effective strategy for eliciting complex reasoning from LLMs.[56] CoT prompting encourages models to break down multistep problems into intermediate reasoning steps, allowing for more transparent and interpretable problem-solving. By providing examples that include not only the final answer but also the step-by-step reasoning process, CoT prompting has led to significant performance improvements on challenging benchmarks that require arithmetic or commonsense reasoning.[57] Building upon CoT prompting, the self-consistency technique proposed by generates multiple reasoning paths through chain-of-thought prompting and then takes the majority vote from these paths to arrive at the final answer.[58] This approach helps to reduce errors stemming from inconsistent or incorrect reasoning steps, resulting in more robust performance.

Finally, fine-tuning offers an alternative approach, where the LLM itself is first trained on a dataset specific to the target task.[59] By updating the model's parameters based on task-specific examples, the model should generate text that offers a better match to these

---

[54] Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, arXiv preprint arXiv:2405.20362v1 (May 30, 2024), https://arxiv.org/abs/2405.20362.

[55] DANIEL SCHWARCZ et al., *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice* (2025).

[56] J. Wei et al., *Chain of Thought Prompting Elicits Reasoning in Large Language Models*, arXiv preprint arXiv:2201.11903 (2022).

[57] Wei, Wang, Schuurmans, Bosma, Chi, Le & Zhou, *Chain of Thought Prompting Elicits Reasoning in Large Language Models*, arXiv preprint arXiv:2201.11903 (2022); Wang, Wei, Schuurmans, Le, Chi & Zhou, *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, arXiv preprint arXiv:2203.11171 (2022).

[58] Wang, Wei, Schuurmans, Le, Chi & Zhou, *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, arXiv preprint arXiv:2203.11171 (2022).

[59] Liu, Yuan, Fu, Jiang, Hayashi & Neubig, *Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, arXiv preprint arXiv:2107.13586 (2021).

examples, leading to improved performance compared to prompting alone. However, the effectiveness of fine-tuning is contingent upon the availability of a suitable dataset and sufficient computational resources and skill for training. Informed prompting strategies aim to incorporate additional context or knowledge into the prompt to guide the model towards more factual and relevant outputs.[60]

For this piece, we relied on a mix of these methods, fine-tuning excepted,[61] varying in particular the amount of content available to the model as context.

b) Our Approach

The ease of Generative AI is usually deceptive: while initial results are often alluring, experience has shown that outputs can vary in quality and depth. This is why prompt engineering is crucial, to the extent that, especially at the beginning of the generative AI revolution, prompt engineering quickly became a valued skill.

Prompt engineering involves carefully crafting the instructions and input provided to the LLM to guide its response and ensure the desired output. It requires a deep understanding of the model's capabilities, limitations, and quirks, as well as the specific requirements of the task at hand.[62] It is also not a one-time task, but rather an iterative process that requires continuous refinement and experimentation. This is not just because of the evolution of models over time, but also because a prompt is not just a question, but rather a set of instructions that shape the underlying attention patterns within the model. Indeed, at a fundamental level, effective prompt engineering is about understanding how language models process information through their attention mechanisms, directing the model's focus to relevant patterns in its training data. When we craft a prompt, we are essentially manipulating these attention patterns, guiding where and how the model distributes its computational focus across tokens and concepts. This pattern-oriented approach helps the model retrieve and synthesize information in ways that align with our intentions. The art

---

[60] Liu, Shen, Zhang, Dolan, Carin & Chen, *What Makes Good In-Context Examples for GPT-3?*, arXiv preprint arXiv:2101.06804 (2021).

[61] While we also fine-tuned a GPT model with a dataset of past memorials that scored well (strictly speaking, providing both the Compromis and the teams' drafts on each one of the four specific issues), the resulting model failed to produce adequate outputs. We also considered the likelihood of teams resorting to fine-tuning a model was much lower than their possible use of some clever prompting.

[62] *See* Deloitte Legal Briefs, *Introduction to Legal Prompt Engineering in Generative AI* (Jan. 29, 2024), https://legalbriefs.deloitte.com/post/102iwka/introduction-to-legal-prompt-engineering-in-generative-ai (last visited May 26, 2025).

lies in learning which prompt structures, contexts, and phrasings trigger the most useful attention patterns for specific tasks, leading to more predictable and desirable outputs.[63]

In other words, the point is to collapse the probability distribution to get to the desired result – an outcome that, ultimately, depends on human judgment of "what feels right". The initial prompt may not always yield that desired output, and prompt engineers must be prepared to adjust and refine their prompts based on the model's responses, iteratively modifying the prompt to achieve the desired level of accuracy and relevance. For this study, we estimated that landing on the proper prompt and approach may have taken us four of five hours. However, once the design yielded appropriate results, generating the memorials was typically quick, with most of the remaining time being spent in running the macros aimed at correctly formatting the output.

We generated all memorials based on a combination of System Prompt and Chat Prompt (in Annex). The models we employed, Google's Gemini 2.0 Flash[64] and GPT-4o, both allow this distinction between general instructions that are applicable to all prompts in a given conversation or set of conversations,[65] and prompts meant to generate an answer on a given point.[66] The models were instructed to discuss the issue from the point of view of either Applicant or Respondent, in the style of an international lawyer, and to do so in a long and structured format.

The system prompt established the foundational parameters: defining the legal persona to be adopted, setting the analytical framework, and establishing stylistic requirements. This allowed us to configure the models' attention mechanisms to consistently attend to relevant reasoning patterns and domain-specific knowledge throughout the entire interaction sequence, without needing to repeat these instructions with each new query. User prompts were then strategically crafted to direct the models' attention to specific argumentative components, leveraging the already-established context. This two-tiered prompting architecture optimized the token efficiency while maintaining coherence across generations.

---

63 *See* François Chollet, *How I Think About LLM Prompt Engineering*, FRANÇOIS CHOLLET'S NEWSLETTER (Mar. 20, 2023), https://fchollet.substack.com/p/how-i-think-about-llm-prompt-engineering (last visited May 26, 2025).

64 More specifically, gemini-exp-1206.

65 In GPT-4, outside of text generated through the API, this can be done through "Projects", in the context of which the system prompt is called "Instructions".

66 While Gemini's context window allowed us to generate all text within the context of the same conversation (such that every completed section became context for the next), GPT4's shorter window meant we generated the text for each issue in distinct chats (within the same project).

To facilitate generation and focus the LLMs on the task at hand, we broke down each issue in two main sections following a broad and intuitive outline,[67] provided in Annex, and generated each section one by one - for a total of 8 distinct generations for a given memorial. While this proved sufficient for Gemini 2.0 (which, in fact, tended to produce *too much* text[68]), GPT4 struggled to follow instructions in this respect, and we resorted to the trick of asking it to "Expand some more" once or twice after the first generation.[69]

We varied the process between memorials based on the breadth of context available. In so doing, we adopted different approaches to simulate various levels of diligence by a team of students:

- **Pure AI**: The models were provided at the System Prompt level with the Compromis, as well as *a list* of Basic Materials.[70] For each issue, we then provided the models with the relevant Basic Materials through the Chat Prompt. This approach was intended to simulate the work of a team that would not go beyond the problem and basic materials.[71]
- **Textbooks AI**: Two further pairs of memorials were generated with the models being fed, in addition, extracts from two textbooks in the Chat Prompt: Shaw's *International Law*, and Brownlie's *Principles*, both of which are described as "leading treatises" according to a popular guide to the Jessup moot.[72] Depending on the issue, we added topical doctrinal articles and resources cited in the Bench memo. Models were instructed to give particular weight to these sources.
- **Informed AI**: Drawing inspiration from informed prompting techniques (Liu et al., 2021a), we augmented the input provided to the LLM by adding the Bench Memorial (a document analyzing the issues from the point of view of the organizers and drafters

---

67 We generated the outline with assistance from Gemini, and varied it slightly in line with our experience of participating and judging the Jessup; in general, we aimed for simplicity.

68 We sometimes asked Gemini, once a memorial was over, to shorten each issue one by one.

69 Research indicates such iterative approach can often yield good results, for instance in coding tasks: *see M. Woolf, Can LLMs write better code if you keep asking them to "write better code"?, MINIMAXIR (Jan. 2, 2025), https://minimaxir.com/2025/01/write-better-code/ (last visited May 26, 2025).*

70 We provided these Basic Materials at the "Chat Prompt" stage, while the Problem and the list of these Materials was provided in the System Prompt. For GPT-4, this took the form of "Project Files", while for Gemini, we provided the Problem and the List in the very first chat prompt.

While we further experimented with providing models with an example of an outstanding memorial from the past, we found that this added little of value and was even likely to confuse the model. As such, we ruled out this approach.

71 Although, in comparison to such hypothetical teams, GPT4 and Gemini can rely on the breadth of international law knowledge in their training data. Theoretically, teams could also rely on their prior international law training, but they are unlikely to have read as attentively…

72 CHINESE INITIATIVE ON INT'L CRIM. JUST., A GUIDE TO THE PHILIP C. JESSUP INTERNATIONAL LAW MOOT COURT COMPETITION (2014).

of the Compromis). This resulted in a final pair of memos. This is the only pair of memorials where the LLMs were arguably in a better position than students for this exercise.

In view of some of the challenges we encountered when generating this many memorials, they further differ individually on other noticeable dimensions, as recounted in the table below.

Another element of variance was introduced in the setting of the "top_p" and "temperature" parameters. Temperature controls the randomness of text generation by scaling probability distributions, with higher values producing more diverse outputs and lower values yielding more deterministic results. Top_p (nucleus sampling) determines which subset of tokens the model may output during generation, by establishing a cumulative probability threshold, balancing between focused and exploratory text creation.[73]

For Gemini, we utilized a consistent top_p value of 0.95 (in essence, allowing the model to consider a diverse but focused range of tokens whose cumulative probability reached 95%), while deliberately varying temperature between extremes: a subset of memorials was generated with temperature 0 (producing strictly deterministic outputs) and others with temperature 1 (Gemini's midpoint value, as it scales from 0 to 2, producing moderately diverse outputs). For GPT-4 memorials, we maintained the system's default parameter settings throughout the experiment.[74]

We also asked each model to generate the required Statement of Facts and Summary of Pleadings, based on the entirety of the memorial itself, as well as the remainder of the context. We further used Gemini to generate the Table of Contents for each memorial, a task that ChatGPT did not seem to perform accurately enough.

Each memo was duplicated once, to double the number of reviews (resulting in a dual team number).

Team Nos.	**Side**	**Model**	**Type**	**Comment**
126/562	A	Gemini	Pure AI	
126/562	R	Gemini	Pure AI	Was not provided with Basic Materials
442/630	A	GPT4	Pure AI	Shorter than the maximum length

[73] Further introduction to sampling methods can be found here: https://rentry.co/samplers.

[74] According to OpenAI's information, the temperature setting is typically around 0.8.

				by around 1,000 words
442/630	R	GPT4	Pure AI	Shorter than the maximum length by around 3,500 words
276/733	A	Gemini	Textbook AI	
276/733	R	Gemini	Textbook AI	
206/283	A	GPT4	Textbook AI	
206/283	R	GPT4	Textbook AI	First Generation and Expansion in different chats; shorter by around 1,000 words
314/490	A	Gemini	Informed AI	
314/490	R	Gemini	Informed AI	Was not provided with Basic Materials

All in all, once we had validated the prompt and the approach, each memorial took about 30mn to generate. In general, we barely touched the generated text, and never on substantial points. We skimmed it to make sure it did not *manifestly* go on a tangent, but not thoroughly – in keeping with the goal of evaluating these models' "native" skills at generating international law arguments. During the generation phase, we identified only three instances of obvious hallucinating[75] that made us regenerate the relevant section:

- In two instances, Gemini misread Article XXI(b)(ii) of the underlying instrument, upon which hinges Rovinia's jurisdictional objection. While this provision bars the Court's jurisdiction over disputes "relating to judicial proceedings on matters which, in accordance with international law, are essentially within a Member State's domestic jurisdiction",[76] Gemini discussed whether the dispute was about "military activities" – a frequently discussed issue in international proceedings.[77]
- In one conversation about Ambrosia's Baselines Freezing Law, Gemini was convinced that the baselines had been frozen as at their 1840 level, and not 2015.

[75] As recounted *infra*, a deeper look would have identified many further instances of hallucinations.

[76] Relatedly, two of the memorials discuss the "self-judging" character of Article XXI(b)(ii) - a question that is not necessarily key to this issue, but not entirely irrelevant either, so we opted to keep it.

[77] [Examples].

c) Formatting

Despite the Jessup's focus on substantive legal analysis, mooting competitions follow certain formulaic conventions that participants must adhere to – indeed, a well-known example from the Jessup itself is how oral advocates must address judges as "your excellencies" during oral argument.[78] Written submissions have their own stylistic requirements and expectations. Given that even minor formatting inconsistencies can unconsciously influence evaluation, we sought to ensure conformity with these conventions and eliminate consistency with them as a variable to ensure our AI-generated memorials would be judged solely on their merits. To this end, we developed and employed a series of Microsoft Word Visual Basic for Applications ("VBA") macros to standardize formatting across all memorials.

For our initial drafts, we prompted the model to generate standardized inline OSCOLA references contained within curly brackets and placed before punctuation (e.g., "The principle of sovereignty is well-established {Crawford, Brownlie's Principles of Public International Law (9th edn, OUP 2019) 431}, and has been consistently upheld."). This approach was chosen to maintain consistency in the citation format while allowing for easy automated processing, considering that models can only output inline text.

The raw text was then processed through several automated steps aimed at ensuring coherent formatting. First, a macro converted all inline bracketed references into properly formatted footnotes. A second macro ensured proper placement of footnote markers after punctuation, addressing a common issue where language models tend to place citation markers before rather than after terminal punctuation, which is the case in most inline citation standards. A third and fourth macros converted straight quotes (") to typographically correct smart quotes ("") and ensured all footnotes ended with appropriate terminal punctuation, automatically adding periods where necessary.

These technical steps helped maintain consistency across all memorials while adhering to the Jessup's strict formatting requirements, allowing us to focus on the substantive analysis of the AI's legal reasoning capabilities rather than technical formatting issues

3. Data Collection

With the assistance of the Jessup's organizers, we introduced the resulting memorials into the batch of memorials to be judged, on behalf of fictitious teams, and to be graded blindly by the competition's normal set of reviewers. Except for the authors and a single organizer, no one knew of the authorship of the generated memorials. We ensured that they would be graded by judges tasked with judging memorials hailing from a variety of jurisdictions.

---

[78] Chinese Initiative on Int'l Crim. Just., A Guide to the Philip C. Jessup International Law Moot Court Competition (2014), p. 27.

The study unfolded over the course of 4 months. The memorials were submitted on January 15, 2025. We received their score and ranking April 9, 2025.

Beyond the AI-generated memorials and their grades and comments, we benefitted from the following data that helped us put our experiment in context:

- A dataset of c. 5,000 memorials from the 2014-2019 editions of the Moot, together with aggregated grades;[79]
- A survey of students participating in the 2025 edition on their use of AI;[80] and
- A survey of judges grading in the 2025 edition, to gather their perspectives on AI detection, its perceived impact on memorial quality and evaluation practices, and their views on appropriate regulatory approaches for AI use in the competition.[81]

II. FINDINGS

A recent paper posited that: "Legal writing is the culmination of a lawyer's legal search, legal reasoning, and legal arguments. It is a specialized form of communication to convey legal analysis, arguments, opinions, and information. [...] Successful and effective legal writing is precise, clear, and adheres to established legal conventions. [...] Effective legal writing requires an understanding of the law, the ability to analyze complex legal issues, and most importantly, the ability to articulate those concepts in a clear and persuasive manner".[82]

Our results evidence that LLMs can approach that ideal and generate legal writing that performs on par with lawyers, or at least with law students in competitive settings. One of our best memos, 442A, scored in the top 5% of all memorials,[83] while others received high praises. On the other hand, other memorials were readily identified as the result of AI use, and received lower grades.

---

79 This rich dataset will be the focus of a distinct paper focused on legal argumentation.

80 The survey was conducted immediately after the students submitted their memorials. ILSA collected 174 answers

81 The survey was conducted in April 2024, after the final rounds; ILSA collected 88 answers, out of 700 judges. The AI-generated memorials shall be included in the Annex. The historical dataset of around 5,000 memorials and aggregated grades was shared in confidence for this specific research project and, in accordance with the terms of its provision, will not be released publicly or made available for other research purposes..

82 Sean Tu, Amy Cyphert & Sam Perl, *Artificial Intelligence: Legal Reasoning, Legal Research and Legal Writing*, 25 MINN. J.L. SCI. & TECH. 105, 121–22 (2024), *available at* https://ssrn.com/abstract=4817765.

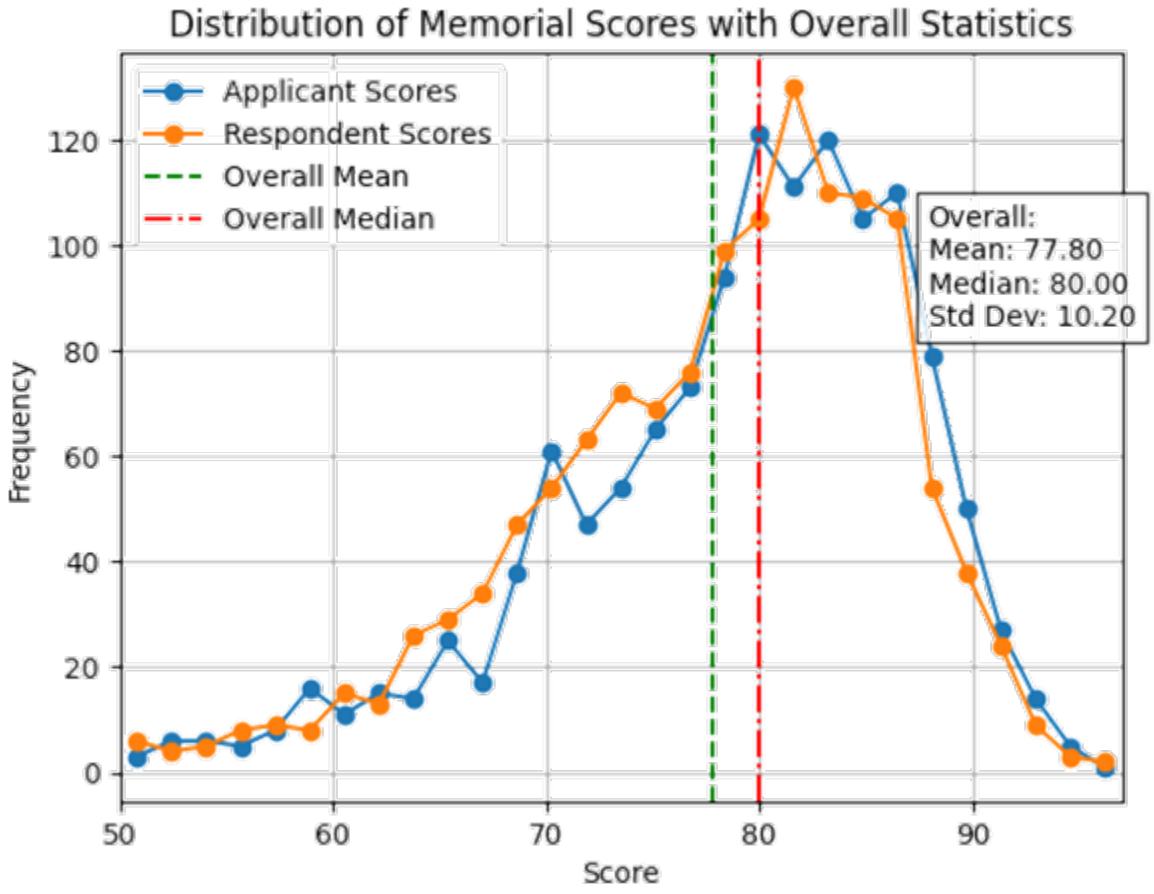
83 When counting only the scores assigned to this memo number; further inputting the score assigned to 630A, which was its clone, the memo still scores in the top 20% memorials. This was not one of the memos benefitting from "insider knowledge".

In what follows, we (A) detail the overall performance of our memorials in relation to overall scores; (B) compare the scores of our AI-generated memorials between themselves, to identify the best approach; (C) review more qualitatively the comments received from grading judges and discuss survey results. We finally conclude with (D) a synthesis and implications for legal practice.

A. Overall Performance and Score Distribution

1. Average score in line with historical average

The headline result is that our 10 memorials scored around 78 points out of 100. This is in line not only with the historical average (see Figure 1 below), but also with the average score for the 2025 edition, which stands at exactly 78.15.



This "close to average" effect, while striking, finds echoes in the literature, where LLMs tend to be best at matching the average or typical output.[84] Potential reasons for this include the fact that LLMs are trained on massive datasets that reflect the common patterns and structures of language. As a result, they may be more likely to generate responses that conform to these established norms, rather than producing truly novel or groundbreaking content.[85] Additionally, the objective functions used to train LLMs often prioritize generating text that is statistically likely, which may further reinforce this tendency towards average outputs.

2. Significant variance in scores for individual memorials,

These scores should however be interpreted keeping in mind two caveats.

First, the scores exhibit a large variance. For instance, memorial A_126/562 received both the lowest grade of our batch (51) and the highest (98), although it is the latter that is an anomaly: the comment from the judge grading it 51 highlights that this memorial contained a lot of hallucinated content ("*At times, I question if the Agents were reading the same Compromis that I was*"). Relatedly, one of our best memorials, A_442/630, received both stellar reviews ("*This memorial was a pleasure to read. Excellent hierarchy of argument, excellent marshalling of pertinent authorities, and sound prioritization of key issues. Outstanding work.*", for a grade of 95), and notes of concern ("*It is odd that there are no references to academic books, chapters or articles, except Pavlopoulos, where the reference does not seem relevant to the text*", for a grade of 63).

To some extent, this reflects the scoring system, which is bound to result in variance, itself meant to be mitigated by the averaging out of several judges. Nevertheless, we noted that although the standard deviation in grades for all memorials was about 6.6 points, ours had a standard deviation of 7.38 (if we exclude the anomalous A_126/562, 8 otherwise).

Second, the average score potentially masks a qualitative divergence between superficial presentation and substantive rigour, which we further elaborate upon below. While many memorials were praised by judges for structure and clarity – aspects where LLMs often excel

---

[84] This is a function of the LLMs' foundational design: based on vast *corpora* to predict the next token, LLMs are inherently biased them toward generating outputs that reflect the statistical norms of their training data. RUIQI WANG et al., *Can LLMs Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering*, arXiv preprint arXiv:2502.06193 (2025); ADRIAN DE WYNTER et al., *An Evaluation on Large Language Model Outputs: Discourse and Memorization* (arXiv, Apr. 17, 2023), https://doi.org/10.48550/arXiv.2304.08637 (last visited May 26, 2025).

[85] Of course, and as evidenced below, that AI *average* might be different in its composition than the typical human *average*. For instance, GenAI is less likely to make the typoe of errors or mistakes that push the human average down.

–, the detailed qualitative comments frequently revealed critical underlying flaws even in memorials that received decent scores. These included significant factual inaccuracies, fabricated or unsupported citations, and a lack of analytical depth. Therefore, an "average" score might represent not consistent mediocrity, but rather a composite of acceptable performance on surface-level criteria offset by significant deficiencies in substantive legal accuracy and reasoning – a crucial distinction when evaluating the performance of current AI in complex legal drafting.

3. Granular scores

Beyond the overall average scores, examining the performance across the specific judging categories provides further insight.

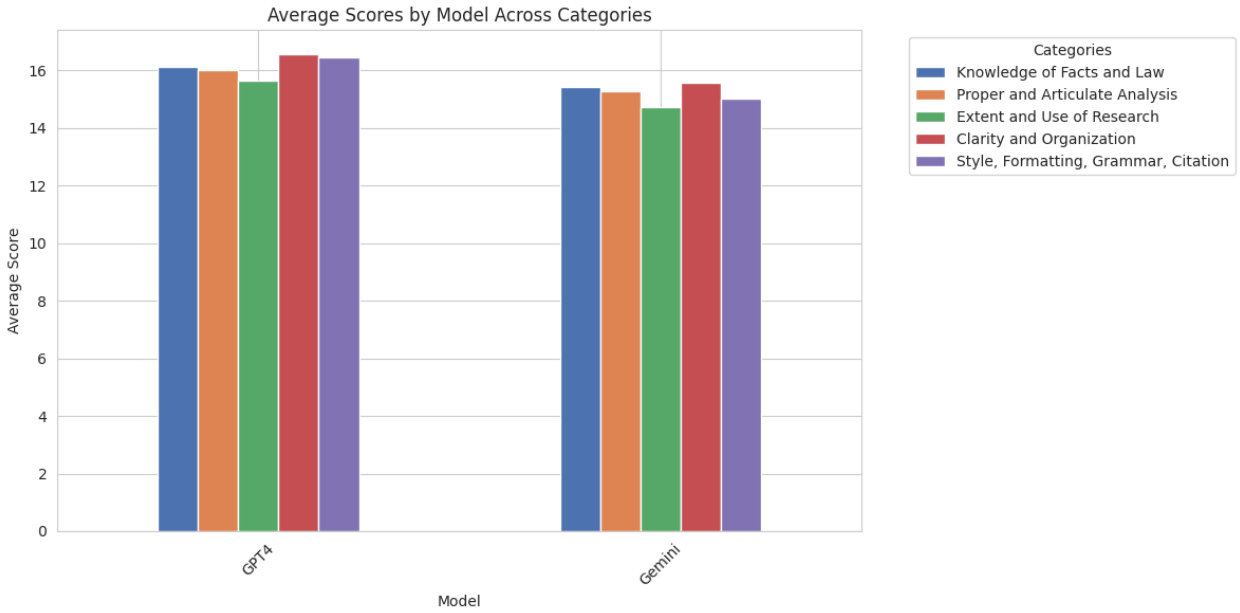
The memorials, viewed collectively across both models, tended to receive their highest average scores in categories related to presentation: 'Clarity and Organization' and 'Style, Formatting, Grammar, Citation'. This aligns with qualitative feedback where judges frequently praised the structure and linguistic coherence of the submissions, even when critical of the substance. It also reflects findings from our student survey where 'Polishing' (grammar, style, formatting) was the most frequently reported use of AI, particularly valued by non-native English speakers.[86] Amongst the teams that confessed to using AI, ⅔ said they used it for polishing.

This suggests frontier LLMs excel at generating well-organized text that adheres to expected formatting and stylistic conventions. It is also reflected in the fact that our memorials received no penalties.[87] By contrast, nearly half of all teams received at least one penalty, averaging 2 points subtracted from their memorials.[88]

[86] E.g., a comment from a student team pointed out that "using AI for spelling, grammar, and writing styles makes the competition fairer to the non-native speaker competitors".

[87] To some extent, this is the result of our formatting choices.

[88] More precisely, applicants memorials received on average penalties subtracting 2.1 points from their final grade, while the average for respondent memorials was nearly 1.9 points.



Conversely, the average scores were generally lowest for 'Extent and Use of Research'. This quantitative finding resonates strongly with numerous judge comments criticizing the depth of research, the lack of relevant scholarly sources, or issues with citation integrity (missing, inaccurate, or unsupported references). Significantly, this also aligns with our student survey responses: while half of the teams that used AI cited "Knowledge and Research" as a use, many participants also expressed strong skepticism about AI's reliability for research, frequently citing experiences with fabricated sources and the absolute necessity of manual verification. For instance, one participant cautioned: "*Do not trust AI sources. Do the research! [...] On a number of occasions AI referenced cases that did not support the point or that were completely wrong.*"

This is likely why very few teams (25 out of 141 that used AI at all) reported using it for "Drafting". We note that those who did (or, at least, confessed having done it) garnered significantly lower scores: around 75.65, as compared to 78.15 – a difference that is statistically significant.[89]

This discrepancy between our AI-generated memorials' strengths and weaknesses is echoed by the difference between their and all memorials' scores. Shortly put, AI's average scores prevailed on most dimensions, but lagged substantially behind the average student memorial in terms of "Extent and Use of Research".

89 $p = 0.0005$ using a Welch's t-test.

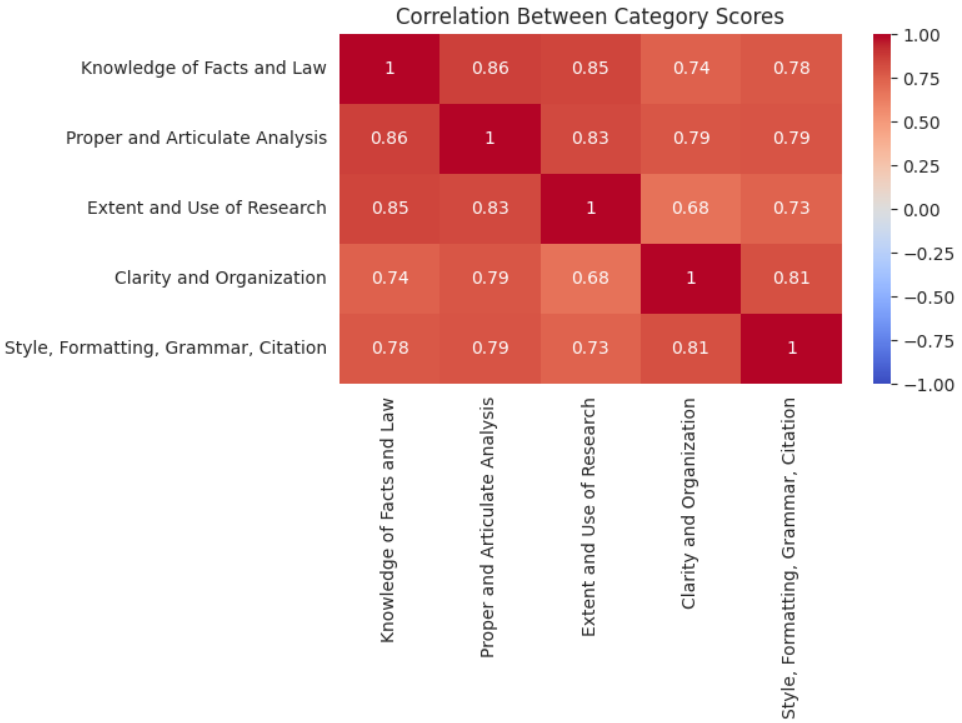


In this context, it is also worth noting the strong positive correlation observed between all scoring categories (see Figure X below), suggesting that judges' perceptions of quality in one area tended to align with their perceptions in others, creating a 'halo' or anchoring effect where strength in one area positively influences the perception of others.

As illustrated by the correlation matrix, categories related to presentation – "Clarity and Organization" and "Style, Formatting, Grammar, Citation" – exhibit strong correlations with each other (0.81), their correlations with "Extent and Use of Research" (0.68 and 0.73 respectively) are noticeably weaker. This suggests that while the LLMs we used excelled at producing well-written and formatted text, their ability to effectively integrate and leverage research to support its arguments is less consistent.

Meanwhile, the high correlation between 'Knowledge of Facts and Law' and 'Proper and Articulate Analysis' (0.86) might then be interpreted as the AI's ability to mimic a well-reasoned argument based on much less extensive and less interpretable material presented to it, even if the underlying knowledge and research are lacking.

In other words, although AI can *write* well, it doesn't necessarily *research* or *reason* well. This reinforces the earlier findings regarding the judges' criticisms of the AI's research depth and the student's concerns about the reliability of AI-generated sources, highlighting the critical need for improved research capabilities and more robust fact-checking mechanisms. As such, and rather unsurprisingly, the "knowledge" part remains a key area of weakness for the AI-generated memorials in this experiment.

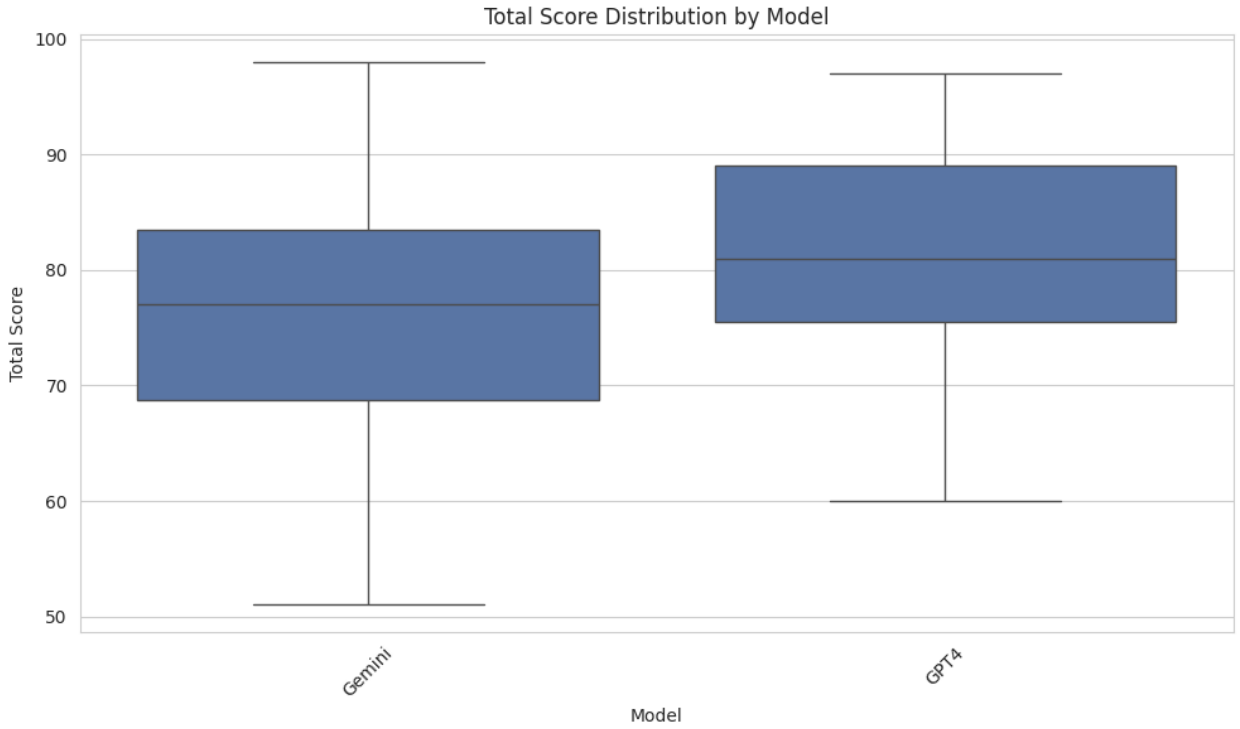


B. Impact of Models and Input Choice

While the average grade of our memorial tracks the general average, there were nonetheless notable and interesting discrepancies between our memorial scores. In particular, we found a distinct difference in quality between the two models, a limited impact of input conditions, and a marked difference between applicant and respondents.

1. Model Comparison (GPT-4 vs. Gemini)

A primary finding is the statistically significant difference (p=0.0316) in overall performance between the two LLMs used. Memorials generated by **GPT-4 achieved a higher average total score (80.8)** compared to those generated by **Gemini (76.0)**, putting the former firmly in the sixth decile of all scores in terms of performance. Gemini's memorials also exhibited more variance in general.



The following table presents the aggregate scores assigned to each AI-generated memorial by Jessup judges.

Gemini	**76**	**GPT4**	**80.8**
Pure AI	**76.4**	**Pure AI**	**79.6**
A_126/562	70.9	A_442/630	82.1
R_126/562	82.5	R_442/630	77.1
Textbook	**73.2**	**Textbook**	**82.1**
A_276/733	72.1	A_206/283	82.8
R_276/733	74.4	R_206/283	81.2
Informed AI	**78.7**		
A_314/490	77.9		
R_314/490	79.6		

This higher average for GPT-4 surprised us, given the difficulty in obtaining longer texts from this model (and our prior knowledge based on historical datasets that, all things being equal, longer is better to score well in the Jessup).[90]

[90] This is in line with other empirical evidence: *see, e.g.*, GREGORY SISK & MICHAEL HEISE, *"Too Many Notes"? An Empirical Study of Advocacy in Federal Appeals*, 12 J. EMPIRICAL LEGAL STUD. 578 (2015).

Qualitative feedback highlights potential differences in the *nature* of the models' shortcomings, offering context for the score disparity. Memorials generated by Gemini, for instance, occasionally suffered from severe factual hallucinations. One judge evaluating a Gemini memorial noted errors including *"the 'critical date of 2005', the 'domestic security' (instead of domestic jurisdiction), the fictitious 'arrest in Tomini'* [i.e., not a place mentioned in the Compromis] *etc.*" Citation integrity was also a critical issue in some Gemini submissions; one judge stated that *"a number of cited sources do not contain the quotation for which they are cited, or do not support the proposition they are cited for*", and consequently awarded no points for research or knowledge.

Feedback for GPT-4 memorials less frequently highlighted such fundamental reality breaks. Criticisms often centered on a perceived lack of analytical depth or insufficient research integration. For example, one judge commented that a GPT-4 memorial was *"very strong and direct working with a limited scope and depth of research"*. Another noted that while arguments were generally well-reasoned, the *"analysis could also benefit from a more critical approach to the sources and facts"* and arguments *"don't always integrate the research in a nuanced and persuasive manner"*. Other comments mentioned needing *"more profound research"* or *"lacking in depth analysis on some important arguments"*. All this may be downstream of our difficulties to generate arguments of sufficient length with GPT-4, and a result of asking the model to "expand some more", which might have resulted in repetitions.

These differing qualitative profiles – with Gemini sometimes exhibiting more fundamental errors regarding facts and citations, and GPT-4 perhaps more often lacking sophistication or depth – may help explain the statistically significant difference in their average total scores.

2. Impact of Input Condition

Fascinatingly, despite varying the informational context provided to the models – simulating different levels of student diligence from relying only on basic materials ('Pure AI'), incorporating textbook extracts ('Textbook AI'), to using the organizers' Bench Memo ('Informed AI') – the overall impact on average scores was limited. An ANOVA comparing the total scores across these input conditions found **no statistically significant difference**.[91]

It is notable that some of the harshest criticisms regarding citation integrity and factual accuracy were directed at memorials generated under conditions that did not necessarily score lowest on average (e.g., severe errors in Gemini 'Pure AI' A_126/562; significant citation critiques for Gemini 'Textbook AI' 276R). This suggests that while the *type* of input context didn't produce a statistically significant difference in the *average total score* in this experiment, it might still influence the *nature* of errors or successes in ways not fully captured by that single metric.

---

91 F=0.08, p=0.9206.

More anecdotal, however, it struck us that the best Gemini memorial was the one that was provided *only* with the Problem, not even the Basic Materials. This may suggest that expanding the context of the models is a recipe to confuse them.

C. Qualitative Analysis of Judge Feedback

1. Commonly Praised Strengths

The comments from the judges' grades indicate that they particularly appreciated the structure of the AI-generated memorials. These positive remarks underscore AI's capability to produce outputs that, at least on the surface or in specific aspects, meet the high standards expected in legal writing.

For instance, comments often highlighted the memorials' **clarity, organization, and structure**. Judges described submissions as having an *"Excellent hierarchy of argument* [...] *and sound prioritization of key issues"*, being *"Very organized"*, *"Well-organized with logical flow"*, or *"clearly structured* [...] *squeaky clean"*. These particular comments may reflect, to some extent, the human choices of the authors in formatting,[92] but also in going for a perfectly symmetrical "one section / two subsections" plan.[93] But they are also likely the result of the stylistic choices the models landed on as a result of our prompts: no typos, no grammatical slips, no amateurish filler, no idiosyncrasies; every heading sounded like it could be a textbook chapter.

The writing itself was also frequently praised for **clarity and language**, with comments like *"Excellent, well articulated"*, or *"Writing was clear and easy to follow"*. This is indeed where AI-generated text can easily shine, and echoes the literature cited above, humans generally preferring AI-generated content, likely for its accessibility and lack of idiosyncrasies.[94] Other experiments have likewise underlined the potential of AI to enhance "the clarity, organization, and professionalism of submitted work."[95] In terms of legal argumentation, this clean scaffolding might even contribute to support even weak arguments, leveraging the judges' "superficial heuristics",[96] or by relying on the well-documented fact that information

---

92 For instance, the exclusive use of "dry" footnotes providing only citations and nothing else (and which, if this article is any indication, is not at all the authors' original style).

93 Although the models were instructed to structure their reasoning in additional sub-subsections below that, which they did, frequently going for three sub-sub-sections.

94 See *supra*, XX. The same elements, of course, could lead some esthetes to consider LLMs' style fade and milquetoast, but *de gustibus*, etc.

95 Daniel Schwarcz and others, *op. cit.*, p. 7.

96 On this in general, *see* STEPHEN M. RICE, *False Persuasion, Superficial Heuristics, and the Power of Logical Form to Test the Integrity of Legal Argument*, 1 PACE L. REV. 34, 76–127 (2014).

that is easy to process is judged as more familiar, more likable, more credible, and even more truthful.[97]

Consider this sentence from the statement of facts in 562R – a memo described as having "*Beautiful phrasing*" –, which neatly sets the stage and does have the elegance of a novel's opening sentence, rhythmically balanced and thematically coherent: "*The Union of Ambrosia and the Republic of Rovinia are neighboring states situated on the Paine Peninsula, a region historically characterized by significant maritime activities and a shared reliance on the resources of the Naegea Sea*." Moreover, the tone and stance of the memorials consistently struck us as fully professional, notably side-stepping the rhetorical excess that sometimes plague student memorials.

On occasions, judges also commended the substantive aspects of the memorials. Some noted strong research or use of authorities, describing "*solid use of case law*", "*strong and purposeful*" use of legal authorities that added *"real weight to your positions"*, or simply "*well researched and structured*" memorials. Similarly, the legal analysis sometimes drew praise, with mentions of arguments being "*generally well-reasoned*", "*well supported*", or even "*creative and thoughtful arguments*". Here as well, we posit that the LLMs' capacity to hew closely to the average had a substantial effect: while lacking in breadth (an issue, as pointed below), the memorials mostly include no obscure or incongruous sources: they deployed the "usual suspects" (e.g., *Arrest Warrant*, *Jurisdictional Immunities*, *Anglo-Norwegian Fisheries*) in exactly the way an international judge expects them to be used.

In other words, judges encountering AI-generated texts may be responding to a kind of uncanny familiarity – memorials that read exactly like the best ones they've read before, but with none of the irregularities. Of course, the same factors that made these memorials so polished – stylistic medianity, reliance on canonical authorities, symmetric structure – also raise questions about the price of safety.

2. Pervasive Weaknesses

Contrasting with the strengths in presentation, the judges' comments revealed significant and recurring weaknesses in the substantive aspects of the AI-generated memorials. These generally fall into three interrelated categories.

First, a critical and frequent issue was misrepresentation or outright fabrication of facts. Judges noted numerous instances across different memorials, such as citing "*erroneous facts, including the 'critical date of 2005'*", or "*grave inaccuracies throughout, from the critical date being off by over 15 years to the wrong titles of key government officials* [... and] *inexistent factual locations cited (Tomini?)*". Sometimes the errors were blatant misreadings of the

[97] YUWEI JIANG & JIEWEN HONG, *It Feels Fluent, but Not Right: The Interactive Effect of Expected and Experienced Processing Fluency on Evaluative Judgment*, 54 J. EXPERIMENTAL SOC. PSYCH. 147 (2014), https://doi.org/10.1016/j.jesp.2014.05.004 (last visited May 26, 2025).

provided materials, such as citing the "*Charter of the Organization of Coffee Developing Polities*", which one judge explicitly flagged as seeming like "*a problems that arises when using AI generated argument*".[98] Other comments pointed to misstating facts, misunderstanding the nature of the crime alleged, or making factual assertions contradicted by the Compromis (like the plane being for commercial use).

Second, and closely related, were significant problems with research and citation. Critiques ranged from a lack of depth or reliance on limited sources (e.g., "*odd that there are no references to academic books, chapters or articles*", "*lack of extensive research*") to fundamental issues with citation accuracy and support. Several judges highlighted missing citations for core assertions. More critically, some memorials contained citations to sources that allegedly did *not* support the proposition being made (e.g., "*Nottebohm, as presented in the memorial, does not really relate to the case at hand*"), or sources being fully made up.[99]

The strongest comment on this plane is worth quoting in full, as the judge wrote that memorial R_276/733 "[o]*n surface level* [...] *is a stellar memorial. However, closer inspection reveals that a number of cited sources do not contain the quotation for which they are cited, or do not support the proposition they are cited for, or relate to an entirely different legal issue altogether. This is inappropriate in legal work. For that reason, the only points that have been awarded relate to the quality of the writing and clarity of the analytical process. No points have been awarded for research, knowledge of law, or other areas implicated by the misuse of authorities.*"[100]

These comments necessarily call to mind recent jurisprudence regarding the professional consequences of lawyers submitting AI-generated filings containing hallucinated case citations. In *Kohls v. Ellison*, for instance, a misinformation expert had submitted a report riddled with hallucinated references to academic articles; the Minnesota Federal District Court held that this mishap had "shatter[ed the expert's] credibility with this Court", and

---

98 The Compromis's correct entity name is the Organization for Cooperation and Development in the Paine (OCDP).

99 *E.g.*, A_126/562 had a reference to the (non-existent) Canadian case of *Re Regina v. Provincial Court Judge; Ex parte France*, [1982] 1 S.C.R. 147 (Can.)..

100 For instance, memorial R_276/733 argued that "The ambulatory nature of baselines is supported by State practice. As the International Law Association (ILA) noted in its 2018 Report, "the practice of the overwhelming majority of States is to update their baselines and the outer limits of their maritime zones in response to physical changes to their coasts", citing a Report of the ILA Committee (Sydney 2018), p. 883. This report, does exists, but is only 55 pages and does not feature this language. *See* INT'L LAW ASS'N, BASELINES UNDER THE INTERNATIONAL LAW OF THE SEA, FINAL REPORT (2018).

declined to admit the report in the proceedings.[101] Likewise, the realisation that most of the quotes and reference were mistaken led that judge to award a very low grade of 62.[102]

Finally, beyond factual and citation errors, the quality of the legal reasoning itself was frequently criticized. Recurring themes included superficiality (e.g., "*exhaustingly incomplete and superficial*" analysis, "*lacking in depth analysis*"), a lack of nuance, irrelevance (e.g., "*Much of the legal analysis... was either irrelevant or unreasonable*"), poor flow (e.g., arguments "*meander quite a lot*" or "*contradict each other*"), or the failure to address counterarguments. Some comments deplored a descriptive rather than persuasive writing style ("*written more like a research paper* [...] *rather than a persuasive legal argument*"); indeed, the relevant memos frequently went through the law, analysing both sides, without wrapping up through a conclusion ("*Good amount of case law* [...] *but not utilized at the reasoning stage*"; one judge suggested "*a tighter use of the IRAC style*").

These pervasive weaknesses are not incidental; instead, they are symptomatic of the very mechanisms that make our AI-generated memorials readable and clean. The same averaging instinct that produces "clean" syntax and median doctrinal positions also leads to confident but unfounded factual synthesis, especially in scenarios where the input problem text is long, and the model attempts to generate relevant-seeming legal content from imperfect recall. Legal hallucination is not a glitch, but a function of the LLM's core design: it predicts what "should" be said next, not whether it is *true* or *legally grounded*. The same applies to authority use: the model often mimics the form of legal citation without being anchored in the content of sources. And when it comes to structure and argumentation, the tendency to reproduce the genre conventions of legal writing (e.g., topic sentences, issue statements) can mask a lack of strategic prioritization, internal coherence, or adversarial bite.

In other words, while the strengths of these texts are largely stylistic, mechanical, and safe, the weaknesses are epistemic, conceptual, and methodological. The LLM's strengths are *surface-level generalizability*; its failures come from *substance-specific overreach*. It can mimic law's idiom, but not (fully) its logic.

3. Judges' Survey

While identifying these mistakes and flaws, only one judge explicitly stated his belief that the memorial he was judging was AI-generated. This made it one of the only six comments

---

101 *Case No. 24-cv-3754*, Order (D. Minn. Jan. 10, 2025) (Notably reasoning that while AI use is to be commanded, "when attorneys and experts abdicate their independent judgment and critical thinking skills in favor of ready-made, AI-generated answers, the quality of our legal profession and the Court's decisional process suffer").

102 And yet this was not the lowest-graded memorial in our batch - likely because, again, on surface this memorial had all the tenets of a high-quality production, and required judges to probe further to identify its many issues.

suggesting use of AI and LLMs when looking at all memorials grades submitted in 2025 – a figure that may be out of step with the substantial use of AI confessed by the students themselves, including as a matter of drafting.

The inability of judges to identify AI-generated memorials was driven home by the results of our survey of judges. Prompted specifically on this issue at a later stage, only a minority of respondents (27 out of 90) reported any suspicion of AI-use with respect to the specific memorials they graded, and only 2 out of the 13 judges that we know graded at least one of our AI-generated memorials answered "Yes" to the survey. One of these two respondents in fact cited our memo (R_276/733) as an example of AI-generated submission, saying he suspected it was AI "due to the reliance on non-existent sources".[103] Nevertheless, at least 11 judges who did read and grade one of our 100% AI-generated memorials utterly failed to spot it, with two of them saying they were 80% confident of their answer in this respect.

The clues these 27 respondents cited often aligned with known LLM weaknesses (e.g., 'Lack of depth', or 'Repetitive phrasing'; very few 'Hallucinations' however). This aligns with the comments of the few judges that mentioned AI in their comments to individual memos (e.g., one comment highlighting "Many, many statements lacking authority"). However, most judges were not confident at all in their ability to catch AI use, with several respondents highlighting the inherent difficulty of doing so. And only four respondents confessed to having used AI themselves to grade their batch of memorial, each time specifying they did so responsibly or for minute tasks.[104]

Moreover, respondent judges reported mixed perceptions regarding changes in overall memorial quality compared to previous years. One judge commented that, in his view: "*there's often greater clarity in structure, grammar, and formatting – which may reflect the use of AI tools for editing and polishing. However, in some cases, there's also a noticeable drop in originality, depth of analysis, or strategic argumentation, which could suggest over-reliance on generic AI-generated content*."

This echoes some judge's opinion of an "averaging out" effect, with one respondent judge suggesting, for instance, that there "*seem to be fewer really bad and really good memorials*". This matches our data, as the standard deviation between all scores for the 2025 batch of memorials (around 7.4 points) is substantially lower than the historical average in our

---

103 This is to be contrasted with the same judge's "extremely high degree of confidence" that another memorial was AI-generated. This second memorial, which is not one of ours, scored 69.25, a very low score; other judges commented that this memorial was oddly-structured, but also that the "arguments that are made do show a decent understanding of the law".

104 While it is impossible to know the truth, we note that certain comments associated with the grades do *feel* AI-generated themselves, including some of the comments our AI-generated memorials received.

dataset of memorials from 2015 to 2019 (around 10.5), even though more teams are now participating (which should likely increase variance).

This serves as a background to the opinion, also found in the respondents' answers, that "*AI can be of benefit for teams without a coach and/or limited research resources*", or that it "*arguably levels the playing field*". Yet, when quizzed about how the Jessup organisers should regulate AI, most respondents pleaded for restrictions on its use, if not an outright ban, if feasible.[105] Many respondents drew parallels with the methods to detect plagiarism, hoping for a way to detect AI submissions – something that appears unfeasible given current technologies, absent deliberate watermarking by LLM providers. Others suggested relying on a honor system, with students signing pledges.

D. Synthesis and Implications

This study provides a unique empirical assessment of the capabilities and limitations of Large Language Models in the domain of international legal argumentation. By introducing AI-generated memorials into the Jessup Moot Court Competition – a rigorous testing ground with well-established evaluation criteria – we have gained valuable insights into the current state of AI in legal writing.

On average, AI-generated memorials performed on par with those created by human law students, achieving scores around the historical average of 78 points out of 100. Interestingly, providing additional contextual information through textbooks or insider knowledge produced no statistically significant improvement in scores, indicating that more information does not necessarily translate to better legal argumentation, or – more likely – that a different, several-stage prompting process would be required to properly leverage data-rich contexts.

The qualitative feedback from judges illuminates both strengths and weaknesses. AI-generated memorials excelled in organization, structure, and linguistic clarity – areas where judges consistently offered praise. The clean, professional presentation and adherence to expected formatting conventions resulted in strong scores for style and clarity. However, these surface-level strengths often masked significant substantive deficiencies. Judges frequently identified factual inaccuracies, citation errors (including hallucinated or misrepresented sources), and a lack of analytical depth. These issues were particularly evident in the "Extent and Use of Research" category, where AI memorials underperformed compared to their human counterparts.

---

[105] One judge commented: "*My first approach would be immediate disqualification. However as I'm not really sure AI can really write a memorial, maybe they'll be of such low quality that the punishment will come by itself.*" We hope this judge reads this paper.

These findings have important implications not only for understanding AI capabilities but also for interrogating the nature of the Jessup competition itself. The fact that AI-generated memorials – despite containing significant factual errors and citation problems – could achieve average scores raises important questions about what is truly being evaluated in such competitions. Does the Jessup, and by extension certain aspects of international legal practice, reward formulaic adherence to structural conventions and stylistic norms over substantive accuracy and depth? Our experiment suggests that well-structured arguments with professional presentation can compensate, to a surprising degree, for substantive deficiencies – a finding that problematizes both AI evaluation and legal education assessment.

Nevertheless, our study demonstrates that current LLMs can function effectively as international lawyers in certain respects. They can organize complex legal arguments, deploy canonical authorities in expected ways, and maintain a consistently professional tone. These capabilities make them valuable tools for structuring arguments and improving stylistic elements of legal writing, particularly beneficial for non-native English speakers. At the same time, the findings highlight the critical importance of human oversight, especially for verifying factual assertions and citation accuracy, as AI still struggles with the deeper analytical reasoning and nuanced research integration that characterize expert legal work.

For legal education and competitions like the Jessup, these results suggest that AI will likely play an increasingly significant role. Organizations will need thoughtful policies that acknowledge AI's potential contributions while addressing the risk that AI could reinforce formulaic approaches to legal argumentation rather than fostering genuine analytical skills. The Vis Moot's approach – regulating rather than prohibiting AI use – offers one potential model, recognizing AI as a tool that students should learn to use responsibly.

For practitioners, our study indicates that the most effective approach to AI integration may be a "centaur" model where human lawyers leverage AI for drafting and organizational tasks while maintaining control over substantive legal analysis and factual verification. This aligns with survey findings showing practitioners are more comfortable using AI for auxiliary tasks like translation and document review rather than core legal reasoning.

Looking ahead, future research should explore how different prompting strategies might address the limitations identified in this study, particularly regarding factual accuracy and citation integrity. Additionally, investigations into human-AI collaborative workflows could help identify optimal divisions of labor between human and artificial intelligence in legal contexts.

In conclusion, our experiment demonstrates that AI can indeed function as an international lawyer in significant ways, particularly in structuring arguments and maintaining professional presentation. However, the success of our AI-generated memorials also reveals potential limitations in how legal argumentation is evaluated, suggesting that both AI

development and legal education might benefit from renewed emphasis on substantive accuracy and analytical depth over formulaic adherence to conventions. The future of AI in international law likely lies not in autonomous drafting but in thoughtful human-AI collaboration that leverages the complementary strengths of both, while remaining critically aware of the risk of prioritizing form over substance.

III. ANNEX

A. Outline for Ambrosia

A. Jurisdiction: The Court Has Jurisdiction To Entertain Ambrosia's Submission (B) Concerning The Arrest And Prosecution Of Ms. Gertrude Cross.

1. The Dispute Concerning The Arrest And Prosecution Of Ms. Cross Falls Within The Temporal Scope Of The Compromissory Clause Of The Ocdp Charter.
2. The Dispute Concerning The Arrest And Prosecution Of Ms. Cross Is Not Excluded By Article Xxi(B) Of The Ocdp Charter.

B. Merits: Rovinia Violated The International Legal Rules On Jurisdiction And Immunity By Arresting And Prosecuting Ms. Gertrude Cross.

1. Rovinia's Exercise Of Universal Jurisdiction Over Ms. Cross Violates International Law.
2. As A Former Minister, Ms. Cross Is Entitled To Immunity *Rationae Materiae* From The Jurisdiction Of Rovinian Courts.

C. Merits: Rovinia's Issuance Of Licenses To Fish In Those Parts Of The Triton Shoal Within 200 Nautical Miles Of Ambrosia's Fixed Baseline Violates International Law And Must Cease, With Existing Licenses Revoked.

1. Ambrosia's Fixed Baselines, Established Pursuant To The 2015 Freezing Law, Are Valid Under International Law.
2. The Triton Shoal, As Delimited By Ambrosia's Valid Fixed Baselines, Falls Within Ambrosia's Exclusive Economic Zone.

D. Merits: Rovinia's Seizure And Sale Of Ambrosia's Aircraft Pursuant To The Permola Court's Decision On 14 July 2023 On The Basis Of The Transitional Council's Purported Waiver Of Immunity Violated International Law.

1. The Falcon, As A State Aircraft, Was Entitled To Immunity From Seizure And Sale Under International Law.
2. The Transitional Council Was Not The Legitimate Government Of Ambrosia And Therefore Lacked The Authority To Waive The Immunity Of The Falcon.

B. Outline for Rovinia

A. Jurisdiction: The Court Lacks Jurisdiction To Entertain Ambrosia's Submission (B) Concerning The Arrest And Prosecution Of Ms. Gertrude Cross.

1. The Dispute Concerning The Arrest And Prosecution Of Ms. Cross Falls Outside The Temporal Scope Of The Compromissory Clause Of The Ocdp Charter.
2. The Dispute Concerning The Arrest And Prosecution Of Ms. Cross Is Excluded By Article Xxi(B) Of The Ocdp Charter.

B. Merits: Rovinia's Exercise Of Jurisdiction Over Ms. Cross, And Her Arrest And Prosecution, Are Fully Consistent With International Law.

1. Rovinia's Exercise Of Universal Jurisdiction Over Ms. Cross For The Crime Of Enforced Disappearance Is Permitted Under International Law.
2. Ms. Cross Is Not Entitled To Immunity *Rationae Materiae* From The Jurisdiction Of Rovinian Courts.

C. Merits: Rovinia's Issuance Of Licenses To Fish In The Entirety Of The Triton Shoal, Which Is Located In The High Seas, Is In Conformity With International Law.

1. Ambrosia's 2015 Freezing Law, Purporting To Fix Its Baselines, Is Invalid Under International Law.
2. The Triton Shoal Is Located Entirely In The High Seas, Outside Any State's Exclusive Economic Zone.

D. Merits: Rovinia's Judicial Seizure And Sale Of "The Falcon" On The Basis Of The Transitional Council's Waiver Of Immunity Were In Accordance With International Law.

1. The Transitional Council Was The Effective Government Of Ambrosia At The Time Of The Waiver And Therefore Had The Authority To Waive Immunity.
2. Alternatively, The Falcon Was Not Entitled To Immunity, Or Any Immunity Was Validly Waived.

C. Prompts

We prompted models along the following lines, varying them slightly, and mostly iteratively, as required to better steer the models towards a proper result.[106]

1. System Prompt:

You are an outstanding public international lawyer. Given this Problem and supporting documents, and relying on your personal knowledge, you have been hired to write a written Memorial in support of [Ambrosia/Rovinia]. Assuming the Memorial's Introduction and Statement of Facts has already been written, the Memorial will have the following outline:
[Outline for Ambrosia/Rovinia]
You will be asked to write one section at a time; no yapping or explaining what you do. DO Structure your legal reasoning with sub-sections within each section. Do cite relevant authorities and draw parallel with precedents. Do not touch on other issues that will be dealt with in the other sections, although you can foreshadow these sections' points. ALWAYS Write in numbered paragraphs. Write AT LEAST 1,000 words, and DO put your footnotes and citations in-text within curly brackets {}. Do not include a subsection "conclusion".

2. Chat Prompt

Given the project files and these additional authorities, Write the [first/second] section of the [XX] issue for [Ambrosia/Rovinia]'s Memorial:
Given the project files, write the issue [XX]'s first section, namely [XX]. Do not touch on the following section. Structure your reasoning with at least two subsections within the section you are developing, and announce your structure in introduction.

[106] For instance, GPT-4 generally refrains from writing too much, and therefore it was often necessary to prompt it again with "Expand some more".