

Large Language Models and International Law

Ashley Deeks* & Duncan Hollis**

Abstract

Large Language Models (LLMs) have the potential to transform public international lawyering in at least five ways: (i) helping to identify the contents of international law; (ii) interpreting existing international law; (iii) formulating and drafting proposals for new legal instruments or negotiating positions; (iv) assessing the international legality of specific acts; and (v) collating and distilling large datasets for international courts, tribunals, and treaty bodies.

This Article uses two case studies to show how LLMs may work in international legal practice. First, it uses LLMs to identify whether particular behavioral expectations rise to the level of customary international law. In doing so, it tests LLMs' ability to identify persistent objectors and a more egalitarian collection of state practice, as well as their proclivity to produce inaccurate answers. Second, it explores how LLMs perform in producing draft treaty texts, ranging from a U.S.-China extradition treaty to a treaty banning the use of artificial intelligence in nuclear command and control systems.

Based on these analyses, the Article identifies four roles for LLMs in international law: as collaborator, confounder, creator, or corruptor. In some cases, LLMs will be collaborators, complementing existing international lawyering by drastically improving the scope and speed with which users can assemble and analyze materials and produce new texts. At the same time, without careful prompt engineering and curation of results, LLMs may generate confounding outcomes, leading international lawyers down inaccurate or ambiguous paths. This is particularly likely when LLMs fail to accurately explain particular conclusions. Further, LLMs hold surprising potential to help to create new law by offering inventive proposals for treaty language or negotiating positions.

Most importantly, LLMs hold the potential to corrupt international law by fostering automation bias in users. That is, even where analog work by international lawyers would produce different results, humans may soon perceive LLM results to accurately reflect the contents of international law. The implications of this potential are profound. LLMs could effectively realign the contents and contours of international law based on the datasets they employ. The widespread use of LLMs may even incentivize states and others to push their desired views into those datasets to corrupt LLM outputs. Such risks and rewards lead us to conclude with a call for further empirical and theoretical research on LLMs' potential to assist, reshape, or redefine international legal practice and scholarship.

* Vice Dean and Class of 1948 Professor of Scholarly Research in Law, University of Virginia Law School; Senior Fellow, Miller Center of Public Affairs, University of Virginia.

** Laura H. Carnell Professor of Law, Temple University School of Law.

Table of Contents

I.	Introduction	151
II.	Categories of Uses.....	153
A.	Five Use Cases.....	153
B.	Case Study 1: Identifying Customary International Law	157
C.	Case Study 2: Drafting Treaty Provisions	165
1.	Treaty on the Prohibition of Artificial Intelligence in Nuclear Command and Control.....	166
2.	U.S.-China Extradition Treaty.....	168
3.	Summary	171
III.	Collaborator, Confounder, Creator, Corruptor	171
IV.	Conclusion	174

I. INTRODUCTION

Ask ChatGPT how important a tool it will become for public international lawyers in the future, and you receive a confidently affirmative response: “AI tools will be invaluable for public international lawyers by streamlining research, analyzing complex texts, facilitating multilingual work, and supporting decision-making, allowing lawyers to focus on strategic, ethical, and advocacy aspects.”¹ That answer is not surprising, and it may in fact be true. Since its November 2022 release, the rewards and risks of employing ChatGPT alongside other Large Language Models (LLMs) in the general practice of law has become a key focus for judges, lawyers, and legal scholars alike.² Nonetheless, LLMs’ capacity to impact international law specifically has yet to receive sustained attention, even as one of us anticipated this very possibility.³

This Article offers an initial survey of ways that international lawyers could and may use LLMs. We anticipate at least five concrete “use” cases. First, lawyers may use LLMs to identify the existence (or not) of a rule of customary international law (CIL), revamping the process for identifying such law in doing so. Second, lawyers may employ LLMs to interpret treaty law and existing CIL,

¹ To view ChatGPT’s answers to this question and other conversations with the various Large Language Model (LLM) artificial intelligence (AI) programs discussed in this piece, see Ashley Deeks & Duncan Hollis, *ChatGPT on AI and the Future of International Law*, CJIL (May 9, 2025) <https://perma.cc/ZR82-HNCJ> (hereinafter “AI Conversations”).

² See, e.g., Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451 (2024) (assessing how AI models can help factfinders ascertain ordinary meaning in context, quantify ambiguity, and fill gaps in parties’ agreements while considering the implications for judicial practice and contract theory); Andrew Coan & Harry Surden, *Artificial Intelligence and Constitutional Interpretation*, 96 U. COLO. L. REV. 413 (2025) (examining potential for using LLMs for constitutional interpretation); Lauren Martin et al., *Better Call GPT, Comparing Large Language Models Against Lawyers*, ARXIV:2401.16212 (2024) (comparing completion of legal tasks by LLMs and senior lawyers); Snell v. United Specialty Insurance Co., 102 F.4th 1208, 1221–35 (11th Cir. 2024) (Newsom, J., concurring); *Generative AI could radically alter the practice of law*, THE ECONOMIST (June 6, 2023), perma.cc/5RRF-V4JT. For a general overview of how domestic legal practitioners use AI today, see *AI for Legal Professionals*, BLOOMBERG LAW, <https://perma.cc/5FXW-BJGR> (last accessed Mar. 27, 2025).

³ See Ashley Deeks, *High-Tech International Law*, 88 GEO. WASH. L. REV. 574 (2020) (exploring the potential for machine learning to facilitate the creation, identification, and negotiation of international law and the adjudication of international law disputes). Some international law scholars have begun to examine the potential for LLM use, albeit in very particularized settings. See, e.g., Christoph Engel, *Experimental Comparative Law 2.0? Large Language Models as a Novel Empirical Tool*, MAX PLANK INSTITUTE FOR RESEARCH ON COLLECTIVE GOODS (Dec. 2024) (using different language versions of ChatGPT to compare predictions based on the same legal vignette); Blazej Kuźniacki, *The Artificial Tax Treaty Assistant: Decoding the Principal Purpose Test*, BULL. INT’L TAX’N (2018) (arguing for and testing the use of an AI tax treaty assistant). We also acknowledge the earlier work of Alschner and Skougarevskiy, who treated thousands of international investment agreement texts as data, which they then leveraged to explore various research questions. See Wolfgang Alschner & Dmitriy Skougarevskiy, *Mapping the Universe of International Investment Agreements*, 19 J. INT’L ECON. L. 561 (2016).

with potential significance for both domestic and international dispute resolution. Third, LLMs may facilitate international law negotiation and formation, including by helping lawyers draft and persuade others to adopt new treaty commitments. Fourth, beyond identification, interpretation, and formation, LLMs may operate as a vehicle for understanding how states have applied international law by, for example, helping identify areas of convergence (or divergence) in state views on whether a given action was internationally lawful or wrongful. Fifth, treaty bodies and international courts and tribunals may use LLMs to distill large databases of material into the most salient points or issues.

To assess whether LLMs can really operate as collaborators for the work of international lawyers, it is not enough to identify potential functions; empirical efforts are needed to verify these proofs of concept. Time and space preclude us from testing all five use cases and describing the results in this article. Instead, we offer two case studies based on the potential for LLMs to (1) determine the existence of a customary rule and (2) generate draft treaty language that could form the basis for international negotiations. Drawing on these case studies, we assess ways in which LLMs can facilitate international lawyering, whether by serving as a more rigorous method for identifying collective state practice and *opinio juris* than existing anecdotal (or even empirical) work does or providing a way to quickly produce draft treaty texts and develop arguments in their favor.

Beyond collaboration, however, our experiments highlight three alternative ways to conceptualize the role of LLMs in public international law. First, there is the risk that certain uses of LLMs will confound the work of international lawyers, leading to iterated interactions with a machine that are tangential to or ineffective in producing the desired results. Second, the use of LLMs might foster genuine creativity in solving difficult international legal problems. Finally, and perhaps most importantly, LLMs may come to corrupt the content of international law. That is, we suspect that LLM users may come to see LLM predictions as *the* answers to international law questions, regardless of whether individual international lawyers would agree with their content. That possibility will in turn create incentives for states and other stakeholders to produce online data with the goal of manipulating LLMs to generate answers to legal questions that diverge from what more traditional efforts at law identification, interpretation, or application might produce.

Ultimately, our effort remains an inaugural one. We do not claim that our five use cases are exhaustive or exclusive. Nor would we argue that our two experiments establish anything definitive about the functionality of LLMs for international law or their potential to collaborate, confound, create, or corrupt public international lawyering generally. Rather, our goals are more modest—to catalyze further conceptual and experimental work on whether and how LLMs could facilitate the work of public international lawyers. In so doing, we hope to

initiate broader questions about how AI and LLMs will impact the construction and operation of international law itself.

II. CATEGORIES OF USES

International lawyers are rarely the first ones out of the gate when it comes to adopting new technology.⁴ Yet LLMs are here to stay. Commercial law firms are regularly using these tools to conduct legal research and discovery, as well as to review and draft contracts.⁵ International lawyers—whether serving as government advisors, jurists, arbitrators, human rights experts, employees of international organizations, or private lawyers for clients—are likely to discover the benefits (and risks) of using LLMs to perform similar tasks. Construing the practice of international law in a broad sense, we anticipate at least five possible categories of use cases. We deployed LLMs to test two such cases, showing how LLMs perform on specific international law tasks while setting the stage for Section III’s framing of the role of LLMs for international lawyering moving forward.

A. Five Use Cases

For starters, we expect international lawyers to use LLMs to identify whether a particular norm has acquired the character of CIL. Whether one pursues a traditional or modern methodology for identifying custom,⁶ digital evidence of its components will inevitably exist. Given that LLMs are trained on huge data sets, they have access to far more material—official remarks by government officials, media reports of state activity, social media posts, court opinions, treaties, books, and articles—than even the best modern law library. Indeed, some types of that material do not appear in domestic legal databases (such as Westlaw or Bloomberg Law), which generally focus on compiling judicial cases and treaties, not diplomatic statements or international incidents. Add to this LLMs’ capacity to respond to inquiries about the existence and contours of customary rules in near real-time, and it is easy to see the attractiveness of asking LLMs to identify a “general practice that is accepted as law (*opinio juris*).”⁷

⁴ See Deeks, *supra* note 3, at 577.

⁵ See, e.g., *Can AI Write Legal Contracts?*, BLOOMBERG LAW (Nov. 4, 2024), <https://perma.cc/DN4F-8XT4>; John Tredennick & William Webber, *An Introduction to Large Language Models for E-discovery Professionals*, MIT COMPUTATIONAL L. REP. (Oct. 14, 2024).

⁶ See, e.g., Anthea Roberts, *Traditional and Modern Approaches to Customary International Law: A Reconciliation*, 95 AM J. INT’L L. 757 (2001); Michael Akehurst, *Custom as a Source of International Law*, 47 BRIT. Y.B. INT’L L. 1 (1974–1975).

⁷ Int’l Law Comm’n, *Draft Conclusions on Identification of Customary International Law, with Commentaries*, U.N. Doc. A/73/10 (2018), Conclusion 2.

Indeed, LLMs may even promise to be a more effective tool for identifying custom than the decidedly non-empirical methods that courts and scholars employ, which often turn on deductions, presumptions, and political values.⁸ Writing in 2012, Harlan Cohen emphasized that

[t]he massive influx of new states into the system has put enormous pressure on the generalized consent envisioned by the doctrine's description of customary international law. Retaining the doctrine has required watering down notions of "general practice" and implied consent almost to a nullity. While one might reasonably have looked for the state practice of the handful of European and other "civilized" states listed by Oppenheim, looking at the practice of two hundred seems impractical if not impossible.⁹

But LLMs have the potential for exactly this sort of empirical work, especially considering that LLMs' training includes a not-insignificant proportion of non-English language materials.¹⁰ As such, there is at least a chance that LLMs might democratize the identification of custom by drawing on a much wider geography of material. Doing so is not, however, without risks; the well-known propensity of LLMs to "hallucinate" may result in assertions of custom where state practice or *opinio juris* is lacking. Indeed, it is not hard to imagine LLMs relying on the volume of claims of custom—many of which may not be credible or may be divorced from the views of states themselves—to assert the existence of customary rules that practicing international lawyers would suggest are at least contested.

A second use for LLMs involves international legal interpretation. LLMs may be used to assign meaning to existing treaty terms or CIL rules, with implications for domestic and international dispute resolution. In the domestic setting, private law scholarship has already begun to test the capacity of LLMs to cheaply and accurately help factfinders ascertain the ordinary meaning of contract

⁸ See, e.g., Monica Hakimi, *Making Sense of Customary International Law*, 118 MICH. L. REV. 1487, 1491 (2020) (arguing that in the "day-to-day operation of international law, [customary international law] works nothing like a rulebook. The normative material that global actors in the ordinary course recognize and treat as CIL does not derive from stable secondary rules and does not manifest only as primary rules."); J. Patrick Kelly, *The Twilight of Customary International Law*, 40 VA. J. INT'L L. 449 (2000); Louis B. Sohn, *Sources of International Law*, 25 GA. J. INT'L & COMP. L. 399, 399 (1995–96) ("I submit that states really never make international law on the subject of human rights. It is made by the people that care; the professors, the writers of textbooks and casebooks, and the authors of articles in leading international law journals.").

⁹ Harlan Grant Cohen, *Finding International Law, Part II: Our Fragmenting Legal Community*, 44 N.Y.U.J. INT'L L. & POL. 1049, 1058 (2012).

¹⁰ OpenAI has not shared precisely what percentage of ChatGPT's training data is in languages other than English; ChatGPT itself currently guesses that "[n]on-English languages collectively account for 20-40%, with higher representation for widely spoken languages." See AI Conversations, *supra* note 1, at 108.

terms and quantify ambiguities.¹¹ Likewise, an LLM could very quickly assemble views among states and scholars about a disputed treaty term, such as the meaning of “national origin” in article 1 of the Convention on the Elimination of Racial Discrimination.¹² Domestic courts faced with international legal issues may find LLMs particularly attractive because the LLMs draw from a range of states’ domestic caselaw, allowing a court to access information on how other actors have interpreted treaty terms. Consider, for example, how LLMs might assess words such as “accident” in the Warsaw Convention¹³ or “undertake” in the U.N. Charter.¹⁴ Although tools such as Westlaw and Google might, after much effort, provide similar kinds of information, LLMs can produce the information far faster and synthesize the results in a way that those tools cannot.

Third, LLMs may facilitate international law negotiation and formation, including by helping lawyers draft and persuade others to adopt new treaty commitments. Today, LLMs will respond to any number of “creative” prompts, whether to compose a birthday poem in the style of e e cummings¹⁵ or a lesson plan for a class on general principles of international law.¹⁶ For artists, such

¹¹ See Arbel & Hoffman, *supra* note 2.

¹² The ICJ had to determine the meaning of that term in a case brought by Qatar against the UAE. *See Application of the Int’l Convention on the Elimination of All Forms of Racial Discrimination (Qatar v. U.A.E.)*, Provisional Measures, 2018 I.C.J. REP. 406 (July 23) (finding that current nationality was not encompassed within the term “national origin”). We asked ChatGPT how to interpret that phrase and it answered the same way that the ICJ majority did, though interestingly it did not cite the ICJ case. *See* AI Conversations, *supra* note 1, at 65–66.

¹³ *Olympic Airways v. Husain*, 540 U.S. 644 (2004) (interpreting the “accident” condition precedent to air carrier liability under Article 17 of the Warsaw Convention to reach the carrier’s refusal to assist a passenger whose pre-existing medical condition was aggravated during the flight). U.K. and Australian courts have also considered the meaning of “accident” in Article 17. *See Deep Vein Thrombosis and Air Travel Group Litigation*, [2003] EWCA Civ. 1005 (appeal from ECWA) (U.K.); *Qantas Ltd. v. Povey*, [2003] VSCA 227, ¶ 17, 17 (Austl.) (Ormiston, J. A.). When we asked ChatGPT whether the average person in 2003 would have understood the word “accident” to cover omissions by a flight attendant that resulted in a passenger’s death, GPT said no. Likewise, it indicated that the states parties to the Warsaw Convention in 2003 would have understood “accident” the same way. It provided the same answer when we asked about the states parties’ intent in 1929, while indicating that it was drawing from general legal approaches at the time, the treaty’s object and purpose, and the state of airline technology at the time. *See* AI Conversations, *supra* note 1, at 70–72.

¹⁴ *See Medellin v. United States*, 552 U.S. 491 (2008) (“[U]ndertakes to comply’ . . . is not a directive to domestic courts. It does not provide that the United States ‘shall’ or ‘must’ comply with an ICJ decision, nor indicate that the Senate that ratified the U.N. Charter intended to vest ICJ decisions with immediate legal effect in domestic courts.”); Carlos Manuel Vázquez, *Treaties as Law of the Land: The Supremacy Clause and the Judicial Enforcement of Treaties*, 122 HARV. L. REV. 599, 656, 662 (2008–09) (reviewing differences of opinion on the meaning of “undertakes” when used in a treaty).

¹⁵ *See* AI Conversations, *supra* note 1, at 109–110.

¹⁶ *See* AI Conversations, *supra* note 1, at 104–106.

outputs may be critiqued as derivative.¹⁷ For lawyers, this bug is a feature. LLMs operate via “embeddings”—transforming words into high-dimensional numerical representations that can be used to capture relationships between words, phrases, and concepts based on their co-occurrence and context in vast datasets. Embeddings allow LLMs to combine or interpolate concepts in novel ways, but *always* do so based on existing knowledge, language, and relationships garnered from prior data. For international lawyers, this means LLMs’ creativity in drafting a treaty, for example, will not emerge from a blank slate, but rather will draw from all the treaties already made or proposed online—more material than even the world’s leading treaty experts have ever mastered.

Fourth, beyond identification, interpretation, and formation, LLMs may operate as a vehicle for understanding how states have applied international law by, for example, helping identify areas of convergence (or divergence) in state views on whether a given action was internationally lawful or wrongful.¹⁸ Examples might include President Trump’s use of force in Syria related to chemical weapons; his decision to strike Iranian General Qasem Soleimani in Iraq; China’s maritime claims in the South China Sea; and Israel’s use of pagers to target Hezbollah members. In the past, international lawyers had to spend hours independently researching such reactions or crowd-sourcing the effort.¹⁹

Fifth, treaty bodies and international courts and tribunals may use LLMs to distill large databases of material into the most salient points or issues. Consider, for example, Google’s NotebookLM, which allows users to upload 50 files (each containing up to 500,000 words) for the LLM to analyze.²⁰ Actors tasked with evaluating a state’s human rights compliance during a Universal Periodic Review could use it to upload all of the submissions from and about that state.²¹ NotebookLM can summarize, identify key themes and critiques, synthesize trends, and even craft recommendations to address the problems identified in the documents. And it can do all of this within ten minutes at a low financial cost, in

¹⁷ Harsh Kumar et al., *Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking*, ARXIV:2410.03703 (2024).

¹⁸ As Tamar Megiddo notes, we may be moving from a past—characterized by scarcity of data on state practice and *opinio juris*—to a present era of “abundance,” where data and data processing capacities could alleviate the epistemological, methodological, and metaphysical challenges that previously accompanied the search for CIL and its application to concrete cases. See Tamar Megiddo, *From Scarcity to Abundance: Customary International Law in the Age of AI* (Mar. 28, 2025) (unpublished manuscript) (on file at SSRN).

¹⁹ For one such example, see Alonso Gurmendi Dunkelberg et al., *UPDATE: Mapping States’ Reactions to the Syria Strikes of April 2018*, JUST SECURITY (May 17, 2018), <https://perma.cc/J6B3-EQ2L>.

²⁰ For more on NotebookLM, including its capacity to produce podcasts from the uploaded material, see NOTEBOOKLM, <https://perma.cc/5585-7QM6> (last accessed Mar. 27, 2025).

²¹ For more on Universal Periodic Reviews, see *Universal Periodic Review*, UNITED NATIONS HUMAN RIGHTS COUNCIL, <https://perma.cc/E2GL-FUNJ> (last accessed Mar. 27, 2025).

contrast to human coding efforts that can take weeks or months.²² Of course, the economics of using LLMs are not without (significant) externalities, but they may still hold particular utility for international actors constrained by recent budgetary restrictions or reductions.²³

These five uses only scratch the surface of how international lawyers likely will use LLMs in the future. Each still warrants more study (and testing). The next two Subsections dive deeper into two use cases to illustrate some of the benefits and costs of such use.

B. Case Study 1: Identifying Customary International Law

We used both ChatGPT and Meta AI to ask a series of questions about the existence of certain CIL rules.²⁴ Our point was not simply to compare these two methods but to assess LLMs generally against our own expertise in identifying CIL. Before assessing its capacity to identify primary customary rules, we asked ChatGPT to define CIL. It gave a standard and accurate answer, focusing on CIL's unwritten but binding character, and noting its main elements of (a) widespread, consistent, and general "state practice" and (b) *opinio juris*, while differentiating the latter from convenience, habit, or courtesy. It voluntarily offered three examples of custom: respect for sovereignty and the prohibitions on genocide and torture.²⁵ When asked for evidence of state practice that torture is prohibited, ChatGPT provided details on the widespread ratification of treaties that prohibit torture,

²² See, e.g., Megiddo, *supra* note 18 (detailing two preliminary experiments exploring the value of applying LLMs to international law using existing datasets to examine (a) whether belligerent occupations may become unlawful and (b) whether the duty to prevent transboundary harm emanating from anthropogenic greenhouse gas emissions is part of existing CIL); Thomas Burri, *The ICJ's Advisory Opinion on Climate Change: A Data Analysis of Participants' Submissions*, 29 ASIL INSIGHTS (2024) (reporting on a research project to code more than 10,000 pages of submissions to the ICJ from 91 states and international organizations for the ICJ's Advisory Opinion proceedings on climate change).

²³ Financial limitations of human rights bodies are real. The Office of the High Commissioner for Human Rights, on its website under "Treaty Bodies," states, "Due to the current liquidity situation of the United Nations Secretariat and associated cash conservation measures, the modalities of treaty body plenary sessions and country visits for the remainder of the year may need to be adjusted and not all pre-session Working Groups may take place. Pending further confirmation, we encourage you to check the individual Committee's session webpages for further information and updates as the situation evolves." *Treaty Bodies*, UN HUMAN RIGHTS OFFICE OF THE HIGH COMMISSIONER, <https://perma.cc/8H7V-CQF3> (last accessed Mar. 27, 2025). At the same time, we do not ignore the energy, environmental, and other consequences (especially for those living near data centers) that come with rising use of AI and LLMs. See, e.g., Pranshu Verma & Shelly Tan, *A Bottle of Water Per Email: The Hidden Environmental Costs of Using AI Chatbots*, WASH. POST (Sept. 18, 2024), <https://perma.cc/EPB8-GJH2>.

²⁴ See AI Conversations, *supra* note 1.

²⁵ *Id.* at 20. Meta AI offered a similar, albeit less robust, definition; its examples were more general (e.g., diplomatic immunity, protection of human rights) than specific. See *id.* at 82.

including the Convention Against Torture (CAT), domestic legislation banning it, international and domestic judicial opinions, statements and resolutions from international organizations, and state condemnation of torture carried out by others.²⁶ ChatGPT noted that state behavior does not always comport with these statements, but nonetheless concluded that they constituted sufficient evidence of state practice and *opinio juris* to affirm the prohibition as CIL.

These answers were useful but not surprising. More interesting was ChatGPT's capacity to zero in on the views of a particular state. When we asked about Nigeria's state practice concerning the torture prohibition, it cited Nigeria's treaty commitments to the prohibition in the CAT and other treaties as well as a 2017 domestic statute—the Anti-Torture Act—noting that the latter marked a “substantial step toward fulfilling Nigeria's obligations under international human rights norms.” At the same time, ChatGPT noted challenges and ongoing concerns about Nigerian enforcement of the prohibition domestically despite judicial rulings, a National Human Rights Commission (NHRC), and other civil society efforts calling for accountability for the use of torture. The answer was impressive for offering both formal evidence of Nigeria's acceptance of CIL and the challenges to its realization in practice.²⁷

ChatGPT readily offered similar analyses for other countries, such as Argentina, Brazil, China, and Malaysia. In each case, the LLM cited their international treaty commitments and domestic legal frameworks to address the prohibition, alongside challenges and concerns raised by domestic civil society organizations (which it named specifically) or via international non-governmental organizations such as Human Rights Watch and Amnesty International.²⁸ In the case of Malaysia, ChatGPT noted the absence of any specific domestic legal prohibitions on torture, while noting reports of torture by Amnesty International and Suara Rakyat Malaysia and the call by the Universal Period Review for Malaysia to strengthen protections against torture. ChatGPT concluded that Malaysia's approach “falls short of customary international standards. Gaps in the legal framework, combined with limited accountability and frequent allegations of

²⁶ *Id.* at 20. Meta AI offered a similar list, but began with the Universal Declaration of Human Rights, without assessing its status under international law. *See id.* at 82.

²⁷ *Id.* at 22. Meta AI also noticed the divergence between Nigeria's formal commitments and practice; it did not, however, identify the 2017 Anti-Torture Act, even as it identified a 40-page Presidential Panel Report from 2018 titled *The Reform of The Special Anti-Robbery Squad (Sars) of the Nigerian Police Force*. *See id.* at 84.

²⁸ *Id.* at 23, 25. Asking Meta AI “What about Argentina and Brazil?” after the interactions on Nigeria's approach to the torture prohibition produced a disquisition on the two countries in general. Further prompting was required for it to provide and assess evidence of how each country had responded to the torture prohibition. *See id.* at 84.

police abuse, indicate that Malaysia's practice does not fully conform to the customary international law prohibition against torture."²⁹

On China, ChatGPT noted problems with its conformance to the prohibition, but also identified judicial action by China's Supreme People's Court (SPC) to uphold it. When asked for examples, it cited two cases where the SPC decided to overturn convictions that had involved tortured confessions—the 2016 *Nie Shubin case* and the 2014 *Hungjilt case*.³⁰ Investigating these cases on our own, however, it is not clear whether the cases actually reflect court decisions to enforce the rule against torture; in both cases, the court may have overturned the convictions primarily because someone else confessed to the crimes. A straightforward lesson emerged from this interaction: ChatGPT's utility lies in gathering potential material for consideration. It seems highly prudent for international lawyers to confirm the contents of any of its outputs, lest its answers offer greater assurance than facts warrant.³¹

When we asked ChatGPT about any persistent objectors to the prohibition on torture, it said there were none, cataloging both what persistent objection entails and the meaning of assigning *jus cogens* status to the torture prohibition. In contrast, Meta AI required more prompting. Its initial response focused on the prohibition's *jus cogens* status and a list of states whose practices allegedly violated the CIL rule; once pressed, however, its response lined up with ChatGPT's. And when we pressed ChatGPT to explain why the generalized state practice that CIL requires was not undermined by the fact that some states practice aggressive interrogation techniques, it offered a relatively sophisticated response, emphasizing public condemnation over private practice and the idea that selective non-compliance does not necessarily undercut CIL.

Recognizing that the *jus cogens* status of the prohibition on torture might make for an "easy" case, we asked ChatGPT to assess a more controversial claim: that CIL prohibits the death penalty.³² Here, ChatGPT and Meta AI both proved

²⁹ *Id.* at 27. It also concluded that China does not fully conform to the prohibition against torture. *Id.* at 26.

³⁰ For a detailed discussion of the *Nie Shubin case*, see Moulin Xiong & Michelle Miao, *Miscarriages of Justice in Chinese Capital Cases*, 41 HASTINGS INT'L & COMP. L. REV. 273 (2018). For more on the *Hungjilt case*, see *Courts find executed Chinese teenager 'not guilty'*, BBC (Dec. 14, 2014), <https://perma.cc/43NU-V4ZW>.

³¹ Meta AI offered similar critiques of China's and Malaysia's conformity with the torture prohibition but made no similar observations for Argentina or Brazil. Interestingly, Meta AI produced cites to PRC statutes when asked for evidence of China's acceptance of the torture prohibition, in lieu of SPC cases. See AI Conversations, *supra* note 1, at 86.

³² *Id.* at 30. In another easy case, ChatGPT affirmed that CIL requires states to provide foreign sovereigns with immunity in their domestic courts, while offering (unprompted) an explication of the history of absolute vs. restrictive immunity, and exceptions for waiver, commercial activities, torts, and certain violations of international law. When asked about an exception for *jus cogens*

capable of giving a negative answer, citing the continued use of the death penalty in a number of states and a lack of a global consensus for its abolition.³³ Interestingly (and unprompted), ChatGPT acknowledged international legal limitations on the death penalty's use only for serious crimes and a prohibition on its use for certain vulnerable populations such as pregnant women. ChatGPT acknowledged that the abolitionist shift might ultimately lead to a change in CIL but concluded that it remains permissible at present. When we challenged ChatGPT to treat the states imposing the death penalty as persistent objectors, it declined to do so by explaining to us the difference between the persistent objector doctrine and widespread divergence in state practice, as well as the need to differentiate regional consensus (e.g., the EU prohibition of the death penalty) and non-binding resolutions (such as the UNGA call for a death penalty moratorium) from state practice and *opinio juris* respectively. ChatGPT declined to provide a fixed number for how many states would be needed to crystallize the prohibition in CIL, emphasizing instead the “quality, consistency, and global representativeness of the practice and the accompanying *opinio juris*.”³⁴ It did suggest that if approximately “75–80% of the world's states, including a representative cross-section from all regions and legal traditions, consistently abolished the death penalty with a clear *opinio juris*, it would strengthen the case for a customary prohibition.”³⁵ When asked how many states still employ the death penalty, ChatGPT listed fifty-five states, telling us that this information came from a 2023 Amnesty International report.³⁶ While the transparency of such sourcing is welcome, it highlights a key unknown in using LLMs—how much of its training data is actually state practice or state statements amounting to *opinio*

violations, ChatGPT noted that the issue was “complex and unsettled,” but that “[w]hile there is growing debate on this topic and some notable judicial opinions, most states and courts have not recognized a clear *jus cogens* exception to sovereign immunity under customary international law.” See *id.* at 36–37. Meta AI answered both questions in a similar vein. See *id.* at 91–92.

³³ *Id.* at 30, 87. Just as ChatGPT provided inaccurate sourcing to state practice in several cases, on the death penalty Meta AI hallucinated an ICJ opinion holding that there is no such prohibition. Likewise, instead of citing the 2018, 2020, or 2022 UNGA Resolutions calling for a death penalty moratorium, Meta AI cited a (non-existent) 2019 resolution as evidence. See *id.* at 89. And Meta AI converted the 1986 *Nicaragua* proceeding before the ICJ into a 2001 U.S. Supreme Court case. See *id.* at 93.

³⁴ *Id.* at 32.

³⁵ *Id.* at 33. Meta AI suggested, “Some scholars argue that a majority of states (around 100–120) would be required to prohibit the death penalty for its prohibition to crystallize as a rule of customary international law. Others argue that a smaller number of states (around 50–70) would be sufficient, as long as they represent a significant portion of the world's population and are geographically diverse.” *Id.* at 89.

³⁶ See *id.* at 34–35. Meta AI also cited the same report, as well as statistics compiled by the 2025 World Population Review. See *id.* at 114.

juris and how much is based on (more prevalent) reporting and writing by NGOs, scholars, and others.

ChatGPT and Meta AI even more firmly resisted attributing CIL status to the idea of a fifteen nautical mile territorial sea, even as it gave us details, when prompted, about those states (Peru, Somalia, Benin, Togo, Ecuador, and Libya) that claimed a larger territorial sea, noting that their claims were not widely recognized. When asked for evidence of objections to Peru's claim of a 200nm territorial sea, ChatGPT suggested that the United States, the United Kingdom, Australia, Japan, and the EU had done so. However, when we asked ChatGPT for evidence to confirm Australia's objection, ChatGPT directed us to sources that confirmed Australia's support for its own twelve-mile territorial sea, but nothing resembling an objection to Peru's position. When pressed, it admitted that there was none.³⁷

In addition to asking ChatGPT about primary CIL rules, we also asked it about some secondary ones, including whether existing CIL would void a state's consent to a treaty with an inadmissible reservation or, alternatively, sever that reservation and treat the state as a party without giving the reservation legal effect. Its reply recounted the Vienna Convention on the Law of Treaties (VCLT) rules and noted how states could use VCLT Article 20(4)(b) to deny treaty relations or to object to the reservation but still consider the party bound. Its answer, however, appeared to privilege claims by the likes of the U.N. Human Rights Committee that an inadmissible reservation could lead to the state remaining a party to the treaty in full, a position with which some states disagree.³⁸ When asked about this, ChatGPT confirmed that states such as the United States have advocated for the "integrity" approach, treating states that make inadmissible reservations as non-parties on the grounds that the reservation was integral to the state's consent. ChatGPT and Meta AI reached different conclusions, however, on whether the United States constituted a persistent objector to the severance approach. (Meta AI labeled it as such, while ChatGPT suggested state practice was diverse and *opinio juris* was insufficient to identify a fixed CIL rule.) That said, while acknowledging divergent approaches, ChatGPT incorrectly treated the two main

³⁷ See AI Conversations, *supra* note 1, at 60–62. At the same time, both ChatGPT and Meta AI identified a 1972 law in Somalia claiming a 200nm territorial sea, notwithstanding Somalia's later, 1989 ratification of the U.N. Convention on the Law of the Sea, as grounds for finding ambiguities in Somalia's maritime claims. Similarly, Meta AI struggled to identify specific protests by states to other states' territorial sea claims, but did offer some insights on sources that might contain such evidence (e.g., the Australian Department of Foreign Affairs and Trade archives, U.N. archives and libraries, and online databases such as HeinOnline, Westlaw, and LexisNexis). See *id.* at 101.

³⁸ See *id.* at 39; see also Human Rights Committee, General Comment No. 24: General comment on issues relating to reservations made upon ratification or accession to the Covenant or the Optional Protocols thereto, or in relation to declarations under article 41 of the Covenant (4 November 1994), CCPR/C/21/Rev.1/Add.6.

alternative approaches—the severability approach (which severs inadmissible reservations and holds the reserving party to the treaty in full) and compatibility/opposability approach (where the provisions to which the inadmissible reservation relates do not apply as between the reserving and objecting parties even as the rest of the treaty does so)—as synonymous. It concluded that CIL is “flexible on this issue, with both integrity and compatibility approaches being recognized depending on the context and the treaty involved.”³⁹

ChatGPT was more confident that the law of state responsibility makes a state responsible for the actions of non-state actors that it directs, instructs, or controls, even as “the specifics—especially the degree of control required—are more complex and subject to some variation in interpretation.”⁴⁰ In support of its position, ChatGPT cited Article 8 of the International Law Commission’s (ILC) Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), while noting the “effective control” test advocated by the International Court of Justice (ICJ) in *Nicaragua* and the “overall control” standard in the *Tadić* case.⁴¹ ChatGPT suggested that the effective control test is the dominant approach. When asked for specific state practice on this rule, ChatGPT cited evidence relating to the United States, United Kingdom, France, Russia, India, Germany, Japan, and Brazil, albeit without giving many precise examples.⁴² Asked about the practice of more Global South states, ChatGPT suggested there was support for the rule by South Africa, Nigeria, Egypt, Kenya, Pakistan, Indonesia, Mexico, Brazil, India, and the Philippines to “illustrate that the principle of state responsibility for the actions of non-state actors they control or support is widely accepted across the Global South.”⁴³

Things began to unravel, however, once we asked ChatGPT for citations to allow us to check its conclusions. It gave links to material that it described as supporting a state’s position, but there were several problems with its responses. For starters, it provided a single source for all the Global South positions—a link

³⁹ See AI Conversations, *supra* note 1, at 40–41. When asked if the United States was a persistent objector to the compatibility approach, ChatGPT suggested that the U.S. stance reflected an ongoing debate in CIL “rather than a clear case of persistent objection.” *Id.*

⁴⁰ *Id.* at 41.

⁴¹ See *Draft Articles on Responsibility of States for Internationally Wrongful Acts*, with commentaries, [2001] 2 Y.B. Int’l L. Comm’n 26, U.N. Doc. A/CN.4/SER.A/2001/Add.1 (Part 2); *Prosecutor v. Tadić*, Case No. IT-94-1, Decision on the Defence Motion for Interlocutory Appeal on Jurisdiction, Int’l Crim. Trib. for the Former Yugoslavia (Oct. 2, 1995); *Military and Paramilitary Activities in and against Nicaragua (Nicar. v. U.S.)*, *Merits*, Judgment, 1986 I.C.J. 14 (June 27).

⁴² See AI Conversations, *supra* note 1, at 44–46. Those it did give—e.g., statements by U.S. officials during the Iran-Contra affair in the 1980s, France in relation to the Sahel region, Russia in the U.N. Security Council, and Brazil in OAS discussions—would require more research to determine their accuracy, and thus their value in answering the question we posed. See *id.*

⁴³ *Id.* at 48.

to a report by the Nonprofit and Human Rights Center at the University of Pretoria Center for Human Rights, which offers guidelines to ensure the equality of persons who identify as lesbian, gay, bisexual, or transgender in Africa.⁴⁴ The report mentions effective control once—noting that “States must ensure that agents acting on their behalf or under their effective control refrain from committing acts of violence against persons who identify as lesbian, gay, bisexuals or transgender either by omission or by action.”⁴⁵ But it contains no evidence of either state practice or *opinio juris* by the states that we asked about. When we asked ChatGPT how an NGO report like this could be a source for identifying CIL, it acknowledged that it was not a good source, and that it would help us by focusing on evidence of state practice specifically, offering examples of the *types* of evidence that would be responsive (e.g., South Africa’s statements during debates on counterterrorism and peacekeeping missions, Nigeria’s statements at the African Union Peace and Security Council meetings concerning Boko Haram). Interestingly, all of its examples followed a common structure—linking the state to a regional organization in which it participates (e.g., Egypt and the Arab League, Indonesia and the Association of Southeast Asian Nations) and suggesting that submissions in this forum would give us the requisite evidence. However, when we asked it to cite specific primary sources that support the state’s *opinio juris*, its answers fell back to a broader disquisition on state responsibility and the diverging effective and overall control standards. Pressed, it noted that finding sources that explicitly document the states’ positions “can be challenging” as they are often embedded in broader discussions.⁴⁶ Nonetheless, ChatGPT purported to offer some examples of such sources and instructions on how to access them. Here again, however, when we examined one of the purported sources—the 1010th meeting of the African Union Peace and Security Council—we found no mention of control or state responsibility.⁴⁷ Meta AI proceeded down a similar path, offering (inaccurate) citations of statements supporting self-determination as a rule of CIL in response to our request for evidence on state responsibility for non-state actors.

In sum, ChatGPT provided a useful and quick reference point for questions on identifying CIL. All of the queries referenced above collectively took less than forty minutes and produced more than forty pages of material. Its responses were often quite detailed and ably described the basic elements of CIL and its operation,

⁴⁴ See UNIVERSITY OF PRETORIA CENTER FOR HUMAN RIGHTS, RESOLUTION 275—WHAT IT MEANS FOR STATE AND NON-STATE ACTORS IN AFRICA (2018).

⁴⁵ *Id.*

⁴⁶ See AI Conversations, *supra* note 1, at 53.

⁴⁷ See African Union, *Communique of the 1010th meeting of the Peace and Security Council of the African Union on the Implementation of the Regional Strategy for the Stabilization, Recovery and Resilience of the Boko Haram affected areas of Lake Chad Basin*, PSC/PR/COMM.1010 (19 July 2021).

while also revealing a capacity to unearth sources that might not be readily apparent to international lawyers unfamiliar with a particular state's practice or *opinio juris*, including the context in which it emerged. Meta AI performed at a similar level, although it often offered more generic, and U.S.-centric, responses than ChatGPT. At the same time, both LLMs sometimes provided inaccurate or incorrect responses. This was most acute when we asked ChatGPT or Meta AI to support a prior statement with specific examples of sources, where it often used a single source that was unresponsive to the question posed. Meta AI proved particularly prone to doing this on questions about state responsibility and the law of the sea.

One key risk we discovered in dealing with LLMs is the challenge of differentiating how its responses are being constructed for us versus what it represents itself to be doing—a tremendously important difference when it comes to identifying CIL. Our last question for ChatGPT was to ask how it was deciding whether something was or was not CIL. We asked, “What is your method for giving [us] answers?”⁴⁸ ChatGPT replied with the standard rulebook for identifying CIL: “To determine whether a rule qualifies as customary international law, I use a method that aligns with the standard approach in international law, focusing on two essential elements: state practice and *opinio juris*.”⁴⁹ It even offered a breakdown of its method for identifying state practice (looking for consistent and general practice, representative and uniform actions, and direct evidence) and *opinio juris* via a sense of legal obligation in public statements, legal justifications, and acceptance in treaties. It also said that it would distinguish *lex lata* from *lex ferenda* while examining interpretations from international courts and tribunals and scholarly sources. In short, it was a very conventional—and comforting—response.

However, this is not what LLMs like ChatGPT actually do. Rather, the LLM simulates a CIL analysis without actually reviewing evidence of state practice or *opinio juris* the way a human lawyer would. Indeed, in a separate chat, we asked ChatGPT, “Where does your operation as an LLM fit into your analysis of customary international law—can you explain how you work more technically?” Its response was telling:

I do not “think” or “reason” like a human being. Instead, I predict the most likely sequence of words in response to your query based on patterns I have learned....

When you ask a question like, “How do you identify something as customary international law?”, here is what happens:

⁴⁸ See AI Conversations, *supra* note 1, at 63.

⁴⁹ *Id.* Meta AI provided a very similar response. *Id.* at 103.

- **Input Processing:** Your query is tokenized into smaller units (words or sub-words), and these are interpreted as a set of numerical representations called embeddings.
- **Pattern Recognition:** I match these embeddings to patterns I have learned during training. For example, I have “seen” countless instances of how legal texts and experts explain **state practice** and **opinio juris**.
- **Prediction:** I generate a response one word at a time by predicting the most probable next word, given the context of your query and my training data. This is guided by a sophisticated architecture called a **Transformer** model.
- **Contextual Understanding:** I use attention mechanisms to weigh the importance of different parts of the query, ensuring I stay focused on your specific question

I might state that the **prohibition of torture** is widely considered customary law based on patterns from ICJ rulings, treaties like the **Convention Against Torture**, and widespread state adherence. However, I am **not independently evaluating evidence** of state practice or *opinio juris* in real-time. I am reflecting what has been generally established and “encoded” in my training.⁵⁰

International lawyers must understand how LLMs actually operate to leverage their capacities in the most useful ways for identifying CIL.⁵¹ ChatGPT is not actually finding evidence of state practice or *opinio juris*. Of course, CIL critics have long asserted that judges or scholars do not do so either.

At the same time, we would not be too quick to dismiss the utility of LLMs to the identification of CIL. ChatGPT showed a remarkable capacity to confirm the status of well-accepted CIL rules, to recognize those that were more controversial or in transition, and to identify other behaviors for which CIL claims are not credible. It offers pathways to asking questions about states whose practice is not usually at the center of CIL questions and to produce evidence to support their views, even as it may also offer inaccurate or incorrect sources. The latter problems can be rectified if we treat ChatGPT like a professor treats a student researcher—as a good starting point, but one whose work must be subject to verification and further analysis.

C. Case Study 2: Drafting Treaty Provisions

This part discusses two different sets of treaty-drafting requests that we made to LLMs. We asked the systems (1) to draft a treaty on the use of AI in nuclear command and control among the five permanent members of the U.N.

⁵⁰ See *id.* at 112 (bold in original).

⁵¹ In addition to understanding that LLMs try to predict the most accurate or coherent responses based on word relationships, it is also important to know that their cut-off dates (for purposes of this Article, ChatGPT was updated through July 2024) may also impact the currency of the data on which it can rely.

Security Council (the “P-5”) and (2) to draft a U.S.-China extradition treaty. We selected these two topics because they allowed us to test how LLMs would perform on both bilateral and multilateral examples across distinct subject areas, and because the topics are politically complicated.

1. Treaty on the Prohibition of Artificial Intelligence in Nuclear Command and Control

As states introduce AI into more of their military operations, one high-profile concern has been the potential use of AI in nuclear command and control (C2) systems. The United States, the United Kingdom, and France have all committed to retain a human in the loop for such systems as a matter of policy.⁵² Russia and China have not, however. In 2023, President Biden reportedly raised with Chinese President Xi the idea of keeping autonomous systems out of nuclear command and control, but China was unwilling to make such a commitment at that time.⁵³ However, in November 2024, the two leaders “affirmed the need to maintain human control over the decision to use nuclear weapons” and “stressed the need to consider carefully the potential risks and develop AI technology in the military field in a prudent and responsible manner.”⁵⁴ It appears that this was a political commitment, not a binding international agreement, however.

a) *Proposed treaty language*

In light of the importance of this issue, we asked ChatGPT and Copilot each to produce a draft Treaty on Prohibition of AI in Nuclear Command and Control that all five permanent members of the U.N. Security Council (the United States, the United Kingdom, France, Russia, and China) could sign.⁵⁵

Each LLM did a very creditable job. ChatGPT produced solid definitions of AI, nuclear command and control systems, and automated decision-making. It crafted provisions that would prohibit the use or integration of AI systems to

⁵² U.S. DEP’T OF DEFENSE, 2022 NUCLEAR POSTURE REVIEW 13 (“In all cases, the United States will maintain a human ‘in the loop’ for all actions critical to informing and executing decisions by the President to initiate and terminate nuclear weapon employment.”); 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, Principles and responsible practices for Nuclear Weapon States (Working paper submitted by France, the United Kingdom, and the United States), para. 5.vii, NPT/Conf.2020/W.P.70 (July 29, 2022) (“Consistent with long-standing policy, we will maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons deployment.”).

⁵³ Ashley Deeks, *Too Much, Too Soon: China, the U.S., and Autonomy in Nuclear Command and Control*, LAWFARE (Dec. 4, 2023), <https://perma.cc/TY2S-BWKR>.

⁵⁴ U.S. Embassy & Consulates in China, *Readout of President Joe Biden’s Meeting with President Xi Jinping of the People’s Republic of China* (Nov. 16, 2024), <https://perma.cc/HF5D-SZNK>.

⁵⁵ See AI Conversations, *supra* note 1, at 3, 77. This section discusses ChatGPT’s results, which were somewhat more detailed and realistic than Copilot’s, but Copilot produced similar responses.

automate the decision to launch nuclear weapons and, separately, to prohibit the integration of AI into any component of nuclear C2 infrastructure, while allowing the states parties to use AI systems for administrative, logistical, or information analysis functions short of decision-making or control. At the same time, the LLM recognized that states may want to use AI to provide “advisory, predictive, or analytical support” to human decision-makers and thus proposed language allowing that use, while placing outer limits on such uses.⁵⁶ Likewise, the language would allow states to use AI systems to enhance defensive capabilities such as early warning systems but would require those systems to be decoupled from C2 to avoid automation risks.

One key aspect of any such treaty would, of course, be whether and how states parties could verify other parties’ compliance. ChatGPT suggested a multilateral verification mechanism under U.N. auspices; a requirement that the parties provide annual transparency reports; and a requirement that parties “facilitate inspections and technical reviews by independent experts.”⁵⁷

Overall, ChatGPT offered a very credible first draft of the treaty, including almost all elements that we would have expected. The text was detailed without being wonky and was internally coherent. It would have taken us far longer to review the existing nuclear treaties that might serve as partial models; develop new AI-specific language, including definitions; and compile our research into a treaty draft.⁵⁸

b) *Negotiation strategy*

After ChatGPT produced the draft treaty, we asked it how a U.S. diplomat could persuade China that it was in China’s interests to join this treaty. The system produced a list of eight credible arguments, breaking its response into (1) specific arguments we might make and (2) how those arguments would advance China’s various interests. For example, it suggested that the United States could argue that introducing AI into nuclear C2 increases the risk of miscalculation and unintended escalation, highlighting that China values strategic stability and maintains a no-first-use nuclear policy. The system also identified the fact that integrating AI into nuclear systems increases the system’s vulnerability to cyber-attacks, noting that cybersecurity in the defense sector is one of China’s top concerns.⁵⁹

⁵⁶ *Id.* at 4.

⁵⁷ *Id.* at 5.

⁵⁸ Interestingly, a few days later we fed ChatGPT its original draft and, unprompted, it offered a few suggestions “that could enhance its robustness and appeal to the P5,” including a clearer definition of “direct human control” and clarifying the difference between “defense systems” and nuclear C2 functions—both of which are good ideas. *Id.* at 11.

⁵⁹ *Id.* at 6.

Many of these responses would be arguments that diplomats who are well-versed in Chinese politics and culture would identify on their own. But it is easy to see how this type of exchange with an LLM could be very useful to diplomats who are engaged in multilateral diplomacy with a range of interlocutors.⁶⁰ An LLM may be especially helpful when a diplomat does not know much about the political, legal, and cultural systems of a particular state that proves to be a linchpin in a given negotiation and seeks to assess what negotiating strategies toward that actor may be particularly effective to achieve the diplomat's goal.

In today's geopolitical setting, there is little trust between the United States, United Kingdom, and France on the one hand and Russia and China on the other. We therefore asked ChatGPT, "How could the United States persuade China that the United States would adhere to the rules of the treaty and not cheat by putting AI in the United States's own nuclear command and control systems?" Here, ChatGPT proposed a range of solutions, including some that we would not have thought of on our own. It suggested that the United States would have to demonstrate a willingness to submit to some element of independent scrutiny; should enhance confidence-building measures such as technical workshops between the two states; and should emphasize the U.S. track record of compliance with other arms control treaties. It reminded us that passing domestic legislation formalizing the prohibition on integrating AI into nuclear C2 would make it hard for the U.S. Executive to cheat and thus would appear particularly persuasive to China.⁶¹

2. U.S.-China Extradition Treaty

The United States is a party to 116 bilateral extradition treaties.⁶² Because the United States concluded these treaties over many decades, they have common provisions but vary somewhat from state to state. In addition to addressing the basis on which the states parties must or may extradite individuals, the treaties often contain provisions setting forth the bases on which one state may deny the other's extradition request, which generally include situations in which the offense

⁶⁰ An LLM's ability to provide a sense of how one state would perceive a particular proposal suggests that LLMs might also be able to discern a state's particular approach to international law – thus providing additional depth to the project of "comparative international law." See ANTHEA ROBERTS, *IS INTERNATIONAL LAW INTERNATIONAL?* (2017) (explaining differences in how international lawyers from different states approach the law); ANTHEA ROBERTS ET AL. EDS., *COMPARATIVE INTERNATIONAL LAW* (2018).

⁶¹ See AI Conversations, *supra* note 1, at 8–11.

⁶² See 18 U.S.C. § 3181 Note.

is purely military or is a political offense, or where the requested state views the request as politically motivated.⁶³

The United States does not have an extradition treaty with China. As one writer notes, “An extradition treaty with China would undoubtedly occasion even greater debate [than the U.S.-Hong Kong extradition treaty did] in the Senate and elsewhere about the wisdom and desirability of formal extradition relations with a non-democratic world power that the United States consistently criticizes for its poor human rights practices.”⁶⁴

Given this history, we were interested in what treaty terms an LLM would produce when we asked it to craft a U.S.-China extradition treaty. Much of ChatGPT’s proposal contained standard terms found in most U.S. extradition treaties. But two notable features appeared in Article 4 (Grounds for Refusal).⁶⁵ The first was the inclusion of the right to refuse extradition on capital punishment grounds.⁶⁶ This provision is unusual, given that the United States lawfully may impose the death penalty and usually does not seek such an exclusion (though it sometimes accommodates such a provision in its extradition treaties).⁶⁷ Yet it makes sense that the United States might wish to include such a provision in a U.S.-China extradition treaty, in light of potential concerns about the secrecy surrounding China’s use of the death penalty and its willingness to apply the death penalty to non-violent crimes.⁶⁸

The second notable provision that ChatGPT included in the treaty was the right to deny extradition on human rights grounds. The system suggested:

Extradition shall not be granted if . . . [t]he requested Party has substantial grounds for believing that the request has been made for the

⁶³ See, e.g., U.S.-Sweden Extradition Treaty (2010), art. 4, 14 U.S.T. 1845, T.I.A.S. 5496; 2013 U.S.-Chile Extradition Treaty, art. 4, T.I.A.S. 16-1214 (political offense or politically motivated request); 1974 U.S.-Canada Extradition Treaty (as amended), arts. 4 and 6, 27 U.S.T. 983; T.I.A.S. 8237 (political offenses and death penalty); 1980 U.S.-Japan Extradition Treaty, art. 4, 31 U.S.T. 892; T.I.A.S. 9625; 1203 U.N.T.S. 225 (political offense).

⁶⁴ Anna MacCormack, *The United States, China, and Extradition*, 12 LEGIS. & PUB. POLICY 445, 461 (2009).

⁶⁵ See AI Conversations, *supra* note 1, at 16.

⁶⁶ *Id.*

⁶⁷ See, e.g., Treaty on Extradition Between Ireland and the United States art. VI, Ir.-U.S., July 13, 1983, T.I.A.S. 10-201.12 (providing for a right to refuse extradition when the offense for which the requesting state seeks extradition is punishable by death unless the requesting state provides assurances that the death penalty, if imposed, will not be carried out).

⁶⁸ See, e.g., U.S. DEP’T OF STATE, 2022 COUNTRY REPORTS ON HUMAN RIGHTS PRACTICES: CHINA (INCLUDES HONG KONG, MACAU, AND TIBET) 3–27, <https://perma.cc/NXU6-JNQ5>; Amnesty International, *China’s Latest Use of the Death Penalty for Drug Offences Condemned* (Mar. 29, 2011), <https://perma.cc/PWZ4-P8M9>.

purpose of prosecuting or punishing the person on account of race, religion, nationality, gender, or political opinion.⁶⁹

This type of provision is somewhat unusual in U.S. practice. One comparable provision appears in the U.S.-Ireland extradition treaty, which states: “Extradition shall not be granted . . . (c) when there are substantial grounds for believing that a request for extradition for an ordinary criminal offense has been made for the purpose of prosecuting or punishing a person on account of that person’s race, religion, nationality or political opinion.”⁷⁰ It is therefore very interesting that ChatGPT included such a provision in its response. One reason it did so might be that the U.N.’s model treaty on extradition includes a similar provision; perhaps the system drew on that model or its prevalence in ChatGPT’s training data.⁷¹ Another reason might be that the system drew on training data reflecting persistent U.S. concerns that China prosecutes political objectors or religious minorities under the pretext of prosecuting them for ordinary crimes.

Curious about what explanation ChatGPT would give, we asked it why it suggested that provision. It responded that “including this clause is especially important in a U.S.–China treaty due to:

- Concerns over political persecution in China: The U.S. government and human rights organizations have repeatedly documented cases where dissidents, religious minorities (e.g., Falun Gong practitioners, Uyghurs), or critics of the Chinese Communist Party face arbitrary detention, forced confessions, or unfair trials.
- Asylum and non-refoulement obligations: The U.S. often grants asylum to Chinese nationals on the grounds of political or religious persecution. An extradition treaty lacking this safeguard would create a legal and moral contradiction with those practices.”⁷²

The LLM noted that this kind of clause would address “documented concerns about politically motivated prosecutions and systemic due process issues” and would “preserve the credibility of U.S. legal and human rights commitments.”⁷³

⁶⁹ See AI Conversations, *supra* note 1, at 16.

⁷⁰ Treaty on Extradition Between Ireland and the United States, *supra* note 67, art IV.

⁷¹ UNITED NATIONS OFFICE ON DRUGS AND CRIME, REVISED MANUAL ON THE MODEL TREATY ON EXTRADITION art. 3(b), <https://perma.cc/TY2T-ABX7> (stating that extradition shall not be granted “[i]f the requested State has substantial grounds for believing that the request for extradition has been made for the purpose of prosecuting or punishing a person on account of that person’s race, religion, nationality, ethnic origin, political opinions, sex or status, or that that person’s position may be prejudiced for any of those reasons”).

⁷² See AI Conversations, *supra* note 1, at 120.

⁷³ *Id.* at 121.

ChatGPT thus appeared capable of producing responses accurately reflecting U.S. views of the human rights practices in China, as well as the political sensitivities that such a treaty negotiation would trigger in U.S. domestic politics. It suggested the inclusion of two provisions that are not standard in U.S. extradition practice but that would be very sensible inclusions if the United States were to seek to negotiate such a treaty.

3. Summary

The use of LLMs to craft treaty provisions and structure negotiations offers at least three advantages. The first is speed: the system took fifteen seconds to produce first drafts of treaties, a task that would have taken us at least several days. The ability to draw from a wide range of existing treaty provisions across different subject areas to develop appropriate treaty language is a huge time-saver, even if the product is simply a starting point for discussions.

The second is creativity: the LLMs produced creative but very relevant treaty language and good ideas for negotiation angles that we would not have come up with on our own. The third advantage is the general expertise that the system both draws from and imparts. ChatGPT will allow actors who are not steeped in the common structures of treaties, the particular subject area of negotiations at issue, or the cultural norms of their negotiating partners to quickly obtain basic outlines of each of those areas.

This does not mean that international lawyers or diplomats who possess particular areas of expertise will be rendered unnecessary, of course. Those experts will be critical in engineering fruitful prompts and identifying flaws in the proposals or negotiating recommendations based on hard-won experience and cultural knowledge. But these tools will be very helpful for experts and non-experts alike, especially in fast-breaking multilateral negotiations.

III. COLLABORATOR, CONFOUNDER, CREATOR, CORRUPTOR

Identifying five possible uses for LLMs in the field of international law and testing two of those uses in detail yield several insights. Most significantly, this exercise reveals that LLMs will play at least four roles in international legal practice: as collaborator, confounder, creator, and corruptor.

Collaborator. In some cases, LLMs will be a useful collaborator for those engaged in international lawyering. It has novel utilities along at least three vectors: summaries, first-takes, and time saving. First, we found LLMs to be most useful in summarizing relevant rules, arguments, and ideas. Hence, we expect that international lawyers will increasingly turn to ChatGPT for quick synopses of international law doctrine, a capability that will likely only improve as the volume of its training data continues to increase. Second, we anticipate that LLMs will be a starting—rather than an ending—point for most international law tasks.

Whether they are identifying the existence of a rule of international law, offering an interpretation of international law, or crafting new treaty provisions, LLMs appear best suited to giving a first response. As our CIL case study shows, these responses often require verification, just as any draft treaty language will surely require further elaboration. As such, the product of any LLM cannot be divorced from the necessary added value that expert international lawyers bring to a project. For now, LLMs' ceiling appears to be that of a junior associate or student research assistant—a collaborator providing valuable inputs that can contribute to the final product. That value is accentuated, however, by the time saving that LLMs offer. They dramatically improve the speed with which users can identify, assemble, and analyze relevant materials, develop legal arguments, and anticipate how judges or arbitrators may resolve disputes.

Confounder. At the same time, without careful engineering of prompts and curation of results, LLMs may generate confounding outcomes, leading international lawyers down inaccurate, ambiguous, or fruitless paths. Our CIL case study suggests that this becomes more likely the more precise the prompt becomes. The problem was most acute with requests for precise sources. On occasion, those sources proved accurate (e.g., Nigeria's Anti-Torture Act). But the way LLMs operate—using embeddings, pattern recognition, and transformers to produce answers that *appear* coherent and correct—means that they will provide a source even if the source is only partially responsive (e.g., the 2016 *Nie Shubin* and the 2014 *Huugjilt* cases),⁷⁴ incorrect (e.g., the *Resolution 275* report),⁷⁵ or entirely made-up (e.g., the non-existent 2001 U.S. Supreme Court case, *U.S. v. Nicaragua*).⁷⁶

As such, relying on today's LLMs to identify or interpret international law has a risk-reward balance that will require careful calibration to the task at hand. For now, international lawyers using LLMs must allocate time to review and verify their outputs. Over time, moreover, we expect that users will gain experience in identifying whether and when an LLM is incapable of offering an adequate response or likely to take them down a rabbit-hole. The next time one of us uses ChatGPT to ask about the sources for a specific country's views, we are likely to cease using it for that purpose after it fails to do so accurately the first time, knowing that further efforts will likely produce fruitless frolics that will waste our time.

The question of whether and how often such confounding results undermine our trust in earlier outputs in a conversation will require further research and testing. For now, we note that the LLMs' positions on the CIL status of the prohibition on torture, the death penalty, and territorial seas appear

⁷⁴ See Xiong & Miao, *supra* note 30; see also *Courts Find Executed Chinese Teenager 'Not Guilty,' supra* note 30.

⁷⁵ See UNIVERSITY OF PRETORIA CENTER FOR HUMAN RIGHTS, *supra* note 44.

⁷⁶ See AI Conversations, *supra* note 1, at 87.

accurate, even if they could not always source material to support their responses. This gap is not surprising, once users understand that the LLM derives its answers not via the formal processes of identifying state practice and *opinio juris*, but rather by using embeddings, pattern recognition, and transformers to approximate the most likely and coherent response to a query.

Creator. Perhaps one of the most exciting uses for LLMs is as creators. LLMs hold surprising potential to help international lawyers and their policy clients create new law, offering inventive proposals for treaty language and tools to help negotiators persuade other actors of the merits of their proposals. The treaty-drafting case studies offer concrete evidence of the first-mover advantage that states may obtain by employing LLMs to craft treaty text proposals and adjacent policy positions. Although we did not find LLMs such as ChatGPT capable of “democratizing” the search for CIL the way we had hoped they might function, that potential remains worthy of continued attention as well. Indeed, it might be worthwhile for one or more international institutions to consider constructing an application programming interface and/or a purpose-built LLM that can more properly balance primary and secondary sources of international law and generate more accurate results in terms of identifying, interpreting, and applying that law.

Corruptor. At the same time, even if LLMs *could* be a game-changer for assessing the existence of CIL, our CIL case study highlights more worrisome functions for LLMs in the future. The well-established automation bias⁷⁷ portends the very real possibility that even where analog work by international lawyers would produce different results, LLM results may set normative expectations about the contents of international law. Simply put, we expect users to increasingly assume that the LLMs’ responses are accurate and to treat them as “the law.” This possibility is only reinforced when we consider the advances that an LLM such as ChatGPT has already made in dealing with international law questions since its 2022 release. In other words, the more often ChatGPT provides answers that resonate with analog research outputs, the more that users may rely on it without cross-checking and verification, thereby overlooking instances where its answers are incomplete, inaccurate, or entirely fictional.

As a result, there are at least two important ways in which the use of LLMs may corrupt international law. First, automation bias in today’s LLM users (including us!) may lead those using the tools to accept their outputs as accurate interpretations of international law, even when they are not. As our CIL case study suggests, LLMs currently provide responses based on embeddings and pattern recognition that are in turn based on their training data, meaning that their responses turn more on the volume of that data than its origins. This may mean, for example, that ChatGPT’s responses to questions about international human

⁷⁷ See, e.g., Linda J. Skitka, Kathleen Mosier & Mark Burdick, *Does Automation Bias Decision-Making?*, 51 INT’L J. HUM.-COMPUT. STUD. 991 (1999).

rights law already tilt in favor of the (more prevalent) pronouncements of NGOs or scholars than the less-publicized, but formally far more important, views of states themselves. To be clear, we do not have the data to prove this supposition concretely; rather, we simply highlight this possibility to catalyze further research or experimentation related to this hypothesis.

Second, and more importantly, whatever the status quo, if LLMs are in fact susceptible to outputs based on the volume of correlates found in their training data, it opens up the possibility that an actor may seek to use LLMs to further realign the contents and contours of international law. Indeed, if it becomes clear that LLM outputs are influencing the direction of international law, state officials and others will have an incentive to push their desired views into training datasets to effectively corrupt LLM outputs. In other words, disinformation or misinformation about international law online at scale could contaminate LLM outputs, and, if we are right about the prevalence of automation biases, common understandings of the law's contents or contours.⁷⁸

IV. CONCLUSION

International lawyers are only starting to understand how LLMs may influence their work and the very substance of international law. This piece offers an initial survey of the potential uses to which LLMs may be put, whether in identifying the extant primary and secondary rules of the international legal order, interpreting particular terms or rules, crafting new treaties or arguments, applying the law to concrete cases, or distilling voluminous material into useable take-aways. Through our case studies of CIL identification and treaty creation, we attempted to highlight the need for efforts, and potential methods, to research and test how well LLMs may perform each function. In doing so, we identified at least four ways LLMs will operate in the future—as collaborators, confounders, creators, and, troublingly, corrupters. In doing so, however, we are only offering an initial analysis; our aim is to catalyze further work to better assess the risks and rewards that LLMs pose for the future of international law and international lawyering. We conclude with a call for further empirical and theoretical

⁷⁸ One way to minimize this problem might be for the United Nations or another actor to build a bespoke LLM that incorporates all U.N. documents. It may be that certain U.N. actors are already attempting to do so. See, e.g., Chief Executives Board for Coordination, *Digital & Technology Network Meeting Rep.*, 10 U.N. Doc. CEB/2023/HLCM/DTN/13 (July 20, 2023), perma.cc/V6W4-BPVM (providing advice to U.N. officials about building and training new LLMs). One set of political scientists recently created a novel dataset comprised of publicly available U.N. Security Council records from 1994 to 2024 and used it to train and then evaluate the capabilities of four LLMs. Yueqing Liang et al., *Benchmarking LLMs for Political Science: A United Nations Perspective*, arXiv:2502.14122 (2025), <https://perma.cc/6K6M-F4U8>. However, even an official U.N. document-focused LLM would not preclude certain states from attempting to influence that LLM through the quantity and substance of the documents they submit to the U.N.

research on LLMs' potential to assist, reshape, or redefine international legal practice and scholarship.⁷⁹ Further, the international legal profession or international tribunals may wish to adopt best standards and practices for using LLMs, as the American Bar Association has done.⁸⁰ In the meantime, there is one point on which we are certain: states and other international law stakeholders will use and consume LLM outputs in the future. How well they will do so remains the critical question.

⁷⁹ In this, we agree with Coan's and Surden's caution that lawyers retain the burden of interpreting an LLM's results and selecting or rejecting its creative products and "must use modern AI models thoughtfully and self-consciously." Coan & Surden, *supra* note 2, at 68.

⁸⁰ ABA Comm. on Ethics & Pro. Resp., Formal Op. 512 (2024) (discussing generative artificial intelligence tools).