Working paper

Can technology augment order writing capacity at regulators?

Natasha Aggarwal
Satyavrat Bondre
Amrutha Desikan
Bhavin Patel
Dipyaman Sanyal

www.trustbridge.in

Can technology augment order writing capacity at regulators?

Natasha Aggarwal* Satyavrat Bondre† Amrutha Desikan*
Bhavin Patel* Dipyaman Sanyal† ‡

December 5, 2025

Abstract

This paper critically examines the opportunities and challenges of using technology, in particular Large Language Models (LLMs), to assist regulatory order writing in quasi-judicial settings, with a focus on the Indian context. The paper proposes augmenting rather than replacing human decision-makers, aiming to improve regulatory order writing practice through responsible use of LLMs. It identifies the core principles of administrative law that must be upheld in these settings — such as application of mind, reasoned orders, non-arbitrariness, rules against bias, and transparency — and analyses how inherent limitations of LLMs, including their probabilistic reasoning, opacity, potential for bias, confabulation, and lack of metacognition, may undermine these principles.

The paper reviews international frameworks and case studies from various jurisdictions, highlighting common design principles like human oversight, transparency, non-discrimination, and security. It proposes a comprehensive Problem-Solution-Evaluation (PSE) framework for responsibly integrating LLMs into order writing processes. This framework maps specific technical, design, and systemic solutions to each identified risk, and outlines evaluation strategies — end-to-end, component-wise, human-in-the-loop, and automated — to ensure ongoing alignment with legal standards.

The article concludes with practical recommendations for the development and deployment of LLM-based systems in regulatory environments.

*TrustBridge Rule of Law Foundation (www.trustbridge.in)

†dōnō consulting (www.dono.consulting)

‡The authors would like to thank Renuka Sane, Karan Singh, Prashant Narang, Ajay Shah, Arul Scaria, Rahul Hemrajani, Siddharth Raman, Bhargavi Zaveri-Shah, and participants at the NLS AI and Law Forum 2025, and the XKDR Legal Systems Reform Seminar No. 13, for their comments and suggestions on the draft paper. All errors remain our own.

Contents

1 Introduction

Indian regulators have extensive quasi-judicial powers that they express through adjudicatory orders. As an example, the adjudicatory orders of the Securities and Exchange Board of India (SEBI) can result in the imposition of penalties,[1] de-licencing,[2] debarment,[3] and restraint from accessing the capital markets.[4] Such powers, aside from posing existential threats for regulated entities, also have impacts upon the wider market, and affect market participants' confidence in regulatory institutions.[5] Therefore it is critical that they are exercised in a proportionate, legitimate, and well-reasoned manner.

1.1 The current state of Indian regulatory order writing

We suggest that there is room for improvement as regards the current state of Indian regulatory order writing: such orders are written in inconsistent ways, and often do not withstand challenge in appeal. As an example, SEBI's adjudicatory orders take an inconsistent position on what standard of proof should be applied in insider trading matters: while some SEBI orders cite 'beyond reasonable doubt'[6] as the applicable standard, a majority of orders cite 'preponderance of probabilities'[7] as the relevant standard.

We also note that between 2009 and 2023, the Securities Appellate Tribunal (SAT) allowed, partly allowed, or remanded 52% of the appeals before it that arose from SEBI's orders in insider trading matters.[8] The reasons for the regulator losing ground in appeal before the SAT include a failure to identify an insider correctly, not identifying if the information involved was 'unpublished price-sensitive information', or if there was trading on the basis of such information, all of which are fairly basic requirements for establishing the violation of insider trading.[9]

This problem is not unique to SEBI; in 2023-24, half the appeals preferred against orders of

[1] Sections 11B, 15A - 15I, Securities and Exchange Board of India Act 1992.

[2] See, for example, Section 12(3), Securities and Exchange Board of India Act 1992

[3] Section 11(4), Securities and Exchange Board of India Act 1992.

[4] Ibid.

[5] Natasha Aggarwal and others, "Balancing Power and Accountability: An Evaluation of SEBIs Adjudication of Insider Trading" [2025] Trustbridge Rule of Law Foundation Working Papers ⟨https://trustbridge.in/RePEc/papers/2025_Aggarwaletal_insiderTradingSEBI.pdf⟩.

[6] See, for example: (*Order in the matter of selective disclosure of unpublished price sensitive information by Manappuram Finance Ltd* Securities Exchange Board of India in Order/VV/JR/2019-20/ 7331, 2020)

[7] See, for example: (*Order in the matter of Vision TV Limited* Securities Exchange Board of India in WTM/SM/IVD/ID13/24880/2022-23, 2023)

[8] Aggarwal and others (n 5).

[9] Ibid.

the Competition Commission of India were allowed. Of the 20 appeals that were allowed, 19 resulted in a remand.[10]

A study of appeals from orders of ten state Electricity Regulatory Commissions (ERCs) before the Appellate Tribunal for Electricity (APTEL) shows high proportions of appeals being allowed, partly allowed, or remanded back to the regulator. Some ERCs are worse at defending their decisions than others; among the worst on this measure from the ERCs included in the study, the Maharashtra ERC had its decisions overruled (appeal allowed), overruled in part (appeal partly allowed), or remanded for reconsideration in 158 out of 246 instances, a whopping 64.23% of the time.[11] Such reversals add time and uncertainty to regulatory decision-making in critical sectors like electricity, can have adverse effects on investment decisions, and impact India's overall growth story.

In their study of appeals from orders of the Tamil Nadu ERC before the APTEL, Patel and Sane 2024 show that the regulator loses ground in 52% of appeals. An examination of the reasons for such reversals shows that several could have been avoided if a few fairly basic matters had been addressed: for example, the regulator lost 86% of matters related to questions about the use of its own regulatory powers. This statistic is telling of the quality of regulatory governance, and some improvements in order writing practices, such as thorough examination of applicable precedent, can help avoid such reversals.[12]

These examples indicate poor order writing practices and capacity constraints at the regulator. While appellate tribunals may not be flawless in their evaluation of the quality of regulatory orders, these numbers are striking and definitely point to the need for improvement in regulatory order writing practices.

The requirement of 'application of mind' is central to administrative law concerns about the exercise of discretionary powers by the executive, or delegatees. Orders that do not demonstrate application of mind are liable to be struck down as "an arbitrary exercise of power."[13] While we have not noticed a canonical definition of 'application of mind', a survey

[10] Competition Commission of India, "Annual Report 2023-24" [2024] ⟨https://cci.gov.in/images/annualreport/en/annual-report-2023-241734695318.pdf⟩.

[11] Chitrakshi Jain, Bhavin Patel, and Renuka Sane, "Examining the performance of ERCs at APTEL" [2025] The Leap Blog ⟨https://blog.theleapjournal.org/2025/07/examining-performance-of-ercs-at-aptel.html#gsc.tab=0⟩.

[12] Bhavin Patel and Renuka Sane, "Assessing regulatory capability in Tamil Nadu electricity regulation: Evidence from appeals" [2024] Working Paper 4, Trustbridge Rule of Law Foundation ⟨https://ideas.repec.org/p/bjd/wpaper/4.html⟩.

[13] *Onkar Lal Bajaj v Union of India* (2003) 2 SCC 673).

of decisions of the Supreme Court of India[14] suggests that 'non-application of mind' includes instances where:

1. Statutory requirements are not considered or discussed - such as where SEBI fails to establish the ingredients of a violation while holding someone guilty of having committed that violation,
2. No, or a vague reason is provided for a certain action - such as where identical, vague reasons are provided for imposing different sanctions in similar circumstances,[15] and
3. There is no manifestation of a logical consideration or relevant material and evidence for making a certain administrative decision - such as where an ERC refuses to exercise jurisdiction on a matter without considering the nature of the petition before it.[16]

This suggests that several regulatory orders may already fail the test of 'application of mind', and that there is a need to examine ways in which order writing practices at Indian regulators can be improved. In this context, we ask whether there is scope for the use of technology to do so.

1.2 The growing interest in technology-enabled solutions

Given the current state of affairs, it is reasonable to cast the net for potential solutions as far and wide as possible. It is useful to think of various measures that may help improve state capacity, such as trainings for adjudicators on an ongoing basis. The use of technology may offer implementable answers as well.

A reflection of this is the increased interest in the use of Artificial Intelligence (AI) to resolve procedural inefficiencies at quasi-judicial and judicial authorities. For example, in July 2025,

[14] *Onkar Lal Bajaj v Union of India* (n 13); *Vijay Singh v State of UP* (2012) 5 SCC 242; *Rajat Baran Roy v State of WB* (1999) 4 SCC 235; *SN Chandrashekar v State of Karnataka* (2006) 3 SCC 208; *Mansukhlal Vithaldas Chauhan v State of Gujarat* (1997) 7 SCC 622; *Food Corpn of India v State of Punjab* (2001) 1 SCC 291; *Internet & Mobile Assn of India v RBI* (2020) 10 SCC 274.

[15] Natasha Aggarwal, Bhavin Patel, and Renuka Sane, "The exercise of discretionary powers: The case of debarment and restraint from capital markets" [2024] The Leap Blog ⟨https://blog.theleapjournal.org/2024/07/the-exercise-of-discretionary-powers.html#gsc.tab=0⟩.

[16] Patel and Sane (n 12).

the Kerala High Court adopted a policy for the use of AI tools by the district judiciary.[17] In parallel, there is also increased interest in the use of AI, including Generative AI (GenAI) and Large Language Models (LLMs), in substantive aspects of decision-making and order writing.[18]

On the other hand, some recent events highlight the potential dangers of using such technologies in judicial or quasi-judicial settings in an indiscriminate manner: in December 2024, the Bengaluru bench of the Income Tax Appellate Tribunal (ITAT) passed an order regarding the taxability of a transaction on the basis of non-existent judgments.[19] It is speculated that this order was written using GenAI tools. The order was later withdrawn.[20] Similar concerns arose in March 2025, when the Karnataka High Court ordered a probe against a trial court

[17] See also, for example: Ministry of Law and Justice, "Use of Artificial Intelligence in Supreme Court" (25 July 2025) ⟨https://www.pib.gov.in/PressReleasePage.aspx?PRID=2148356⟩; Ministry of Law and Justice, Government of India, Use of AI in Supreme Court Case Management (20 March 2025) ⟨https://www.pib.gov.in/PressReleasePage.aspx?PRID=2113224⟩; Ministry of Law and Justice, "Digital Transformation of Justice: Integrating AI in India's Judiciary and Law Enforcement" (25 February 2025) ⟨https://www.pib.gov.in/PressReleasePage.aspx?PRID=2106239⟩ accessed 6 August 2025; Team MP, "Boosting judicial efficiency: HC launches AI chatbot, e-bail bond & training centre" [2025] Millennium Post (Published June 28, 2025 at 12:08AM IST) ⟨https://www.millenniumpost.in/bengal/boosting-judicial-efficiency-hc-launches-ai-chatbot-e-bail-bond-training-centre-616646⟩; Nirbhay Thakur, "Justice at your fingertips: How AI is helping Delhi's judges, lawyers deal with caseload" [2025] The Indian Express (Updated May 19, 2025 08:15IST) ⟨https://indianexpress.com/article/cities/delhi/justice-at-your-fingertips-how-ai-is-helping-delhis-judges-lawyers-deal-with-caseload-10014723/⟩; IANS, "AI transforming India's judiciary and law enforcement, making justice accessible to all" [2025] ET CIO (The Economic Times) (Published via IANS) ⟨https://cio.economictimes.indiatimes.com/news/artificial-intelligence/ai-transforming-indias-judiciary-and-law-enforcement-making-justice-accessible-to-all/118570991⟩. Recently, the Securities and Exchange Board of India and the Reserve Bank of India have considered the use of AI by regulated entities. See Securities and Exchange Board of India (SEBI), Consultation Paper on guidelines for responsible usage of AI/ML in Indian Securities Markets (Accessed on August 18, 2025, 2025) ⟨https://www.sebi.org.in/reports-and-statistics/reports/jun-2025/consultation-paper-on-guidelines-for-responsible-usage-of-ai-ml-in-indian-securities-markets%5C_94687.html⟩; Reserve Bank of India, Report of the Committee to develop a Framework for Responsible and Ethical Enablement of Artificial Intelligence (FREE-AI) in the Financial Sector (2025) ⟨https://www.rbi.org.in/Scripts/BS%5C_PressReleaseDisplay.aspx?prid=61018⟩

[18] See, for example: Natasha Aggarwal, Bhavin Patel, and Karan Singh, "A Guide to Writing Good Regulatory Orders" [2025] Trustbridge Rule of Law Foundation Working Papers ⟨https://trustbridge.in/work/a-guide-to-writing-good-regulatory-orders/⟩; Divij Joshi, "Automated Administration: Administrative Law and Algorithmic Decision-Making in India" in *The Philosophy and Law of Information Regulation in India* (Centre for Law and Policy Research 2021) ⟨https://publications.clpr.org.in/the-philosophy-and-law-of-information-regulation-in-india/chapter/automated-administration-administrative-law-and-algorithmic-decision-making-in-india/⟩.

[19] *Buckeye Trust v Principal Commissioner of Income Tax* ITA No1051/Bang/2024.

[20] Shipra Singh, "Did AI hallucination play mischief with a tax tribunal order?" [2025] Mint (Livemint) ⟨https://www.livemint.com/money/personal-finance/chatgpt-artificial-intelligence-ai-itat-bengaluru-bench-tax-buckeye-trust-case-errors-11740544685677.html⟩.

judge who relied on non-existent judgments as precedent to pass an order.[21] Despite such incidents, and the imposition of sanctions by judges and courts, lawyers across jurisdictions continue to rely upon AI-generated non-existent precedent.[22] The Washington Post reported a rising trend of such instances, much to the growing despair of courts.[23] These instances suggest that it would be useful to consider regulating the use of these technologies in legal processes, and to develop ways in which the technology can be used in a safer way to generate more accurate results, rather than attempt to prohibit their use entirely.[24]

Such examples can also lead to extreme positions in the discourse, ranging from those advocating the large-scale adoption of technology to overhaul order writing practices, to those opposing any involvement of AI, and LLMs in particular, in quasi-judicial settings.

We suggest that neither of these extremes is helpful. We acknowledge that there are certain attendant risks to the use of LLMs, and that these need to be addressed. On the other hand, there is a need to recognise that 'order writing' is not a monolithic activity, and that there are various types of orders, as well as several steps involved in the 'order writing' process. It is also important that conversations about the feasibility of using technology in the order writing process be grounded in, or at least account for, the principles of administrative law that apply to regulatory order writing. Some types of orders, and some steps in the order writing process may be more amenable to the use of technological tools than others. It is hard to imagine any strong argument to oppose the automation of procedural orders such as those granting adjournments, but it is easy to imagine why we may be concerned about the probabilistic nature of LLMs giving rise to orders such as those of the Bengaluru ITAT

[21] Moneycontrol News, "Karnataka High Court orders action against trial judge who cited non-existent SC judgments to pass order" [2025] Moneycontrol ⟨https://www.moneycontrol.com/news/india/karnataka-high-court-orders-action-against-trial-judge-who-cited-non-existent-sc-judgments-to-pass-order-12977401.html⟩.

[22] See generally: Devashish Bharuka, "A (Cautious) Case For AI In Legal Research" [2025] LiveLaw ⟨https://www.livelaw.in/articles/a-cautious-case-for-ai-in-legal-research-288044⟩.

[23] Daniel Wu, "Lawyers using AI keep citing fake cases in court. Judges aren't happy." [2025] The Washington Post ⟨https://www.washingtonpost.com/nation/2025/06/03/attorneys-court-ai-hallucinations-judges/⟩.

[24] For example, the Singapore Supreme Court has issued a guide for the use of AI by "court users" (such as prosecutors, lawyers, self-represented persons, or witnesses), emphasising that users assume full responsibility and must fact-check all AI-generated content. Judiciary of Singapore, *Guide on the Use of Generative Artificial Intelligence Tools by Court Users* (Court Guide, Judiciary of Singapore 2024) ⟨https://www.judiciary.gov.sg/docs/default-source/news-and-resources-docs/guide-on-the-use-of-generative-ai-tools-by-court-users.pdf?sfvrsn=3900c814_1⟩.

bench, which had 'confabulations'[25] in it. Similarly, there should be no objection to the use of technological aids to assist precedential research, but there may be valid grounds to oppose letting AI take decisions that affect parties' rights.

With this understanding, we emphasise that our work points to suggestions about how technological tools can augment human decision-making in regulatory adjudicatory processes. We do not suggest that these technological tools replace human decision-making. As such, our eventual recommendations are that such tools be used as aids for processes such as the review of draft orders, rather than to substitute the drafting of such orders by humans entirely. Such a pairing of human decision-making with technological aids can, we suggest, improve the quality of regulatory orders.

1.3 Our contribution and the structure of this paper

The available Indian literature on the topic focuses largely on concerns such as confabulations, accuracy levels, and the over-reliance on technocratic means to improve state capacity.[26] While these concerns are well-identified, there is limited discussion on when and how it may be considered feasible to use LLMs to support quasi-judicial order writing. There is also limited discussion in the literature about how these problems impinge on the principles of Indian administrative law.

Our contribution lies in the integrative work of this paper: we draw upon the design principles articulated in frameworks developed in other jurisdictions and relate them to the applicable principles of Indian administrative law. We use this synthesis to develop a Problem-Solution-Evaluation (PSE) framework that is attentive to international practice, the legal principles underpinning quasi-judicial decision-making in India, and problems and limitations inherent to GenAI and LLMs. We do this in the following manner:

We identify the problems associated with the use of LLMs in quasi-judicial order writing as falling into seven categories: (i) non-application of mind, (ii) the black-box problem, (iii) potential for bias, (iv) the confabulation problem, (v) lack of metacognition, (vi) training corpus, and (vii) data security and privacy. We enquire as to the source of these problems, and what principles of Indian administrative law are impacted by these problems. We do this

[25] An LLM confabulates when it generates output that appears plausible and coherent, but is either factually incorrect or irrelevant to context. In popular discourse, confabulations are often called 'hallucinations', even though the former is more technically appropriate. See, Peiqi Sui and others, "Confabulation: The Surprising Value of Large Language Model Hallucinations" [2024] arXiv ⟨https://arxiv.org/pdf/2406.04175⟩, 2.

[26] See, for example: Joshi (n 18)

by studying judicial decisions that lay down the principles of administrative law governing order writing in quasi-judicial proceedings. We find that the administrative law principles affected by the problems we identified are: (i) the requirement of application of mind, (ii) the requirement of giving reasoned orders, (iii) the requirement of non-arbitrariness, (iv) rules against bias, and (v) the need for transparency. We limit our study to 'order writing', which is one among several parts of the larger adjudicatory process that also includes matters such as how hearings are conducted. As such, we exclude those other parts of the adjudicatory process from the scope of our study, and do not examine the principles of administrative law that govern them.

Next, we examine the order writing process, and the requirements of good order writing. We demonstrate that 'order writing' is a composite term that refers to several interrelated activities. Some of these activities may be more amenable to the adoption of technological aids, and we identify these.

Having identified what aspects of order writing permit the use of technological aids such as LLMs, we sharpen our focus to examining whether LLMs have been used in governmental functions generally, and in judicial and quasi-judicial functions specifically, in other countries. If they have, we examine the primary sources setting out guidelines or frameworks that have been articulated to ensure that their use in such settings avoids undesirable outcomes.

We then study the available literature on the mechanisms of LLMs and methods for their enhancement, including explainable AI techniques, to identify what solutions could be designed to address the problems of administrative law raised by the use of LLMs in quasi-judicial settings. As an example, we try to understand whether these methods can address the administrative law principles of application of mind and reasoned order-giving.

Finally, we combine these suggestions from the literature with the guidelines and frameworks adopted in other jurisdictions to examine how we could design evaluation systems to check the efficacy of the proposed solutions.

This helps us arrive at the 'Problem-Solution-Evaluation' framework that we describe in detail in Section 6. We suggest that this framework should form an integral part of any design for the use of LLMs in quasi-judicial processes. Based on this framework, and our study of the literature, we offer some recommendations for any proposed use of LLMs to assist quasi-judicial order writing. The paper ends with our conclusions.

2 Methods

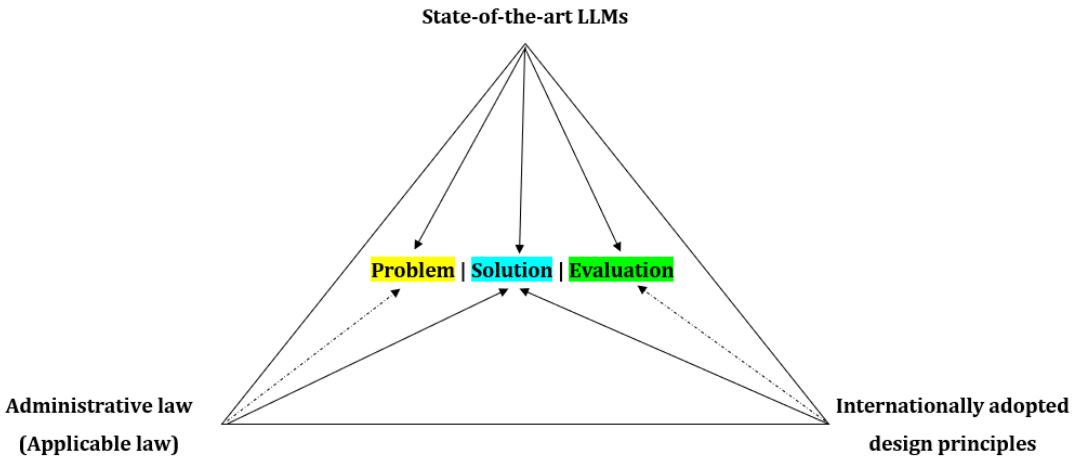
This paper adopts a qualitative research methodology. It relies on a comprehensive literature review to inform the design of a conceptual framework for the use of LLMs for order writing assistance in quasi-judicial settings. To do this, we drew on three distinct domains, namely:

1. **State-of-the-art in LLMs:** We extensively reviewed available scholarly literature on the technical capabilities and limitations of LLMs as they currently stand. From this study, we identified the most frequently occurring problems with LLM-use, as well as potential techniques to mitigate them.
2. **Principles of administrative law:** We looked at the body of Indian administrative law, to identify which of its principles apply to the order writing exercise in quasi-judicial settings, which we define as "Applicable Law".
3. **Internationally adopted design principles:** We conducted a comparative analysis of the legal or policy frameworks that relate to the use of such technologies in judicial or quasi-judicial settings in other jurisdictions. This analysis helped us identify a set of design principles universally considered important for the judicious use of LLMs in high-risk judicial or quasi-judicial environments.

As Figure 1 illustrates, we systematically integrated findings from these three domains to create our conceptual framework. We first mapped the limitations of state-of-the-art LLMs to the requirements of administrative law to define seven key problems. We then devised a set of solutions by mapping commonly adopted international design principles to specific technological and design mechanisms. Finally, informed by the design principles framework, we identified metrics to evaluate these solutions against the benchmarks of administrative law. Together, these components form what we call the 'Problem-Solution-Evaluation' (PSE) framework.

We propose that this interdisciplinary approach is a replicable model for designing a framework for the judicious use of any technology in judicial and quasi-judicial processes. Our work in this paper with LLMs serves as a practical demonstration of this broader application.

Figure 1 This figure represents the three domains which inform the Problem-Solution-Evaluation framework, namely: scholarly literature on the capabilities and limitations of state-of-the-art LLMs, applicable principles of administrative law and internationally adopted design principles.



3 Limitations

We acknowledge the following limitations of this paper:

First, a caveat on the scope of the proposed 'Problem-Solution-Evaluation' framework is in order: We are conscious of the rapid pace of LLM development in both academia and industry. We acknowledge that many of the recommended solutions and evaluation methods discussed in this paper may have evolved significantly by the time this work is public, as may the manner in which LLMs function. Equally, several of the solutions we propose in this paper may be considered contestable or speculative in value. This is also a function of the pace of advancement in LLM development.

Nevertheless, the value of our contribution lies not in the solutions we enumerate as examples, but in the general conceptual framework we devise. We propose that when utilising any technological application to assist with order writing processes, it is crucial to identify possible misalignments with administrative law principles as problems arising from the technological functionality of the application. These misalignments must be highlighted and addressed through meaningful solutions and evaluations. The analysis presented here for an

LLM-supported application serves as an illustration of how this conceptual framework can be practically applied to the responsible development and use of any such application.

Second, we use the term 'Solution' in the 'Problem-Solution-Evaluation' framework in the sense of methods or techniques that (i) help reduce the extent of a problem, (ii) identify where the problem manifests, (iii) provide diagnostic information that can be used to correct errors caused by problems, and (iv) provide insights into how the problem may be removed or ameliorated in later iterations of a technological system. We do not use the term 'Solution' in the sense of a 'cure', which can remove a problem or its symptoms entirely. To illustrate: in the sense in which we use the term, the 'Solutions' to the 'Problem' of hallucination do not make hallucinations go away; rather, they help us: (i) reduce instances of hallucinations, (ii) identify where hallucinations may have occurred, (iii) fix data errors caused by hallucinations, and (iv) devise solutions to reduce instances of hallucinations in further iterations of the system.

Third, our research and suggestions are focused on regulatory order writing, which is one part of the larger adjudicatory process at regulators. Other parts of this larger process include matters such as the procedure for conducting and recording hearings, and the manner in which evidence is received. Other principles of administrative law may apply to these other parts, and we do not consider them in our analysis. Any system that is based on our suggestions would necessarily have to interface with those other steps of the adjudication process, but we do not address them in this paper.

Finally, we acknowledge that the introduction of a set of technological tools will not, by itself, result in any appreciable improvements in regulatory adjudication. Any such tools can be neglected, abused, or misused by users. There is a need to align the incentives of adjudicators at regulatory institutions with the appropriate and intended use of such tools. Intended users must also receive the necessary training to use these tools in safe and appropriate ways. Adjudicators must also be incentivised to provide feedback on the tool, so that it can be iteratively improved and made more useful for them. One way of doing this would be to involve adjudicators from the design stage onwards in any development exercise. This will also require examining current capacity constraints at regulators in detail, redesigning incentive structures, and other behavioural nudges, but a detailed discussion on those is outside the scope of this paper.

4 Problems related to using LLMs in quasi-judicial order writing

LLMs offer promising benefits to improve regulatory capacity in order writing processes. For this, their use in quasi-judicial settings must be compatible with the principles of administrative law that govern these settings. A key challenge arises from limitations inherent to the functioning of LLMs, which often misalign with these principles. In this section, as summarised in Table 1 below, we highlight some limitations inherent to how LLMs work, identify five administrative law principles relevant to order writing in quasi-judicial processes, and demonstrate how the limitations in LLMs can adversely affect these administrative law principles. We also identify two logistical problems that must be addressed in order to integrate LLMs into quasi-judicial order writing processes.

Table 1 An overview of the potential problems that LLMs may raise when used in quasi-judicial order writing, mapped against administrative law principles that they may affect detrimentally.

This table provides an overview of how limitations inherent to LLM functionality may lead to problems that detrimentally affect some principles of administrative law applicable to quasi-judicial processes.

Cause for concern	**Problems**	**Consequence**	**Administrative law principle affected**
Statistical reasoning, not translatable to human (or legal) reasoning - predictive, not deliberative	Non-application of mind	Lack of a sound legal rationale in decision- making processes	Application of mind Reasoned orders Non-arbitrariness
Complex mathematical computations to arrive at output - path of reasoning not traceable/ intelligible to user	Black-box problem	Lack of explainability, auditability of decision- making processes	Reasoned orders Transparency
Replicate biases or toxicity in training dataset or prompt	Potential for bias	Reasonable likelihood of biased output	Rules against bias
Stochastic computational models - predict most likely next word - output may not be accurate or relevant to context	Confabulation	No objective, reasoned, legally sound criteria for decision-making	Non-arbitrariness
No awareness of the limits of available information - confidence-competence gap	Lack of metacognition and the Dunning-Kruger Effect in LLMs	Cannot seek out more information on points of law and fact that they do not have access to	Application of mind Non-arbitrariness
Legal database difficult to collate - legal documents are noisy, unstructured - poor training material	Training corpus	Difficult to build applications, that generate accurate outputs for quasi-judicial contexts	NA
Orders may contain sensitive information - storing information on third-party servers, problematic for interested parties - data breach risks	Data privacy and security	Risk of cybersecurity breaches, detrimental consequences under IP, IT and data protection laws	NA

4.1 Inherent limitations of LLMs

Language models are computational models designed to follow instructions and generate output in natural language. Internally, they are probability distributions over sequences of words (or tokens). LLMs are a type of language model characterised by their immense scale. They typically have billions of parameters in their architecture, and are trained on colossal datasets. This enables them to capture complex linguistic patterns, allowing them to perform a variety of natural language tasks using predictive analysis. While they generate output that appears coherent and at times, even logical or well-reasoned to a user, at their core, LLMs are probabilistic models, engaged in predicting the next likely word based on context.

Our analysis considers four limitations inherent to LLMs' functionality:

1. The quality and nature of training dataset and prompts fed to the LLM at different stages in its life cycle have a direct bearing on the quality of the output they generate. Given their predictive nature, any hidden toxicity in the data, such as biases or inaccuracies, are replicated in the output.
2. The inner workings of LLMs are opaque. They perform complex mathematical computations at massive scales to arrive at their output from a given input. The logical path that LLMs take to arrive at a conclusion is typically indecipherable to the average user. This makes LLMs "black-boxes".
3. Another consequence of the complex statistical reasoning process of LLMs is that their progression from input to output is not comparable or translatable to human logic. It does not imitate the process of deliberation that a human thinker employs. While LLMs can provide output that appears well-structured and coherent to a human, their underlying logic or generative process may not align with generally accepted principles of human thought. This is especially the case in domains that require highly specialised reasoning techniques, such as the legal domain.
4. LLM output is frequently factually incorrect or irrelevant to context.[27] This confabulated content appears coherent and plausible, making it hard to distinguish from accurate or otherwise relevant responses. An LLM user must necessarily therefore treat all output as confabulated, and verify it against a reliable source of truth to avoid the risk of misinformation.

[27] Sourav Banerjee, "LLMs Will Always Hallucinate, and We Need to Live With This" [2024] arXiv ⟨https://arxiv.org/html/2409.05746v1⟩.

A note on post-training alignment and its unintended consequences

In addition to the inherent technological limitations of LLM and the concerns raised by administrative law, the design choices made by developers can also introduce ethical and moral concerns. Post-training alignment serves as an illustration of this. Post-training alignment typically involves techniques like Reinforcement Learning from Human Feedback (RLHF), in which LLM models are fine-tuned to adhere to desired traits such as helpfulness, honesty, and harmlessness. This process embeds normative values into the model, influencing how it generates outputs in response to prompts. A post-training alignment process, which is often referred to as "character alignment", plays a critical role in shaping model behaviour.

This may lead to a problem where the definition of fairness in alignment is predominantly determined by the model provider, guided by the provider's internal policies and ethical guidelines, and societal norms prevalent in the provider's jurisdiction. For instance, providers like OpenAI (developers of GPT models) and Anthropic (developers of Claude) incorporate alignment philosophies that prioritise fairness, equity, and avoidance of harm, often drawing from global human rights frameworks or company-specific values.[a]

This provider-centric approach can lead to misalignments in regulatory settings, in which fairness might be better interpreted through the lens of jurisdictional legal standards, such as those under the Constitution or sector-specific statutes.

Should regulators cede this authority to private entities, or instead establish their own standards for acceptable model behaviour?

We suggest that regulators must take an active role in "defining fairness", perhaps through custom alignment processes or procurement guidelines that require models to be tuned with respect to Indian administrative law principles.

For example, a regulator like SEBI could mandate alignment to ensure outputs prioritise proportionality in penalties, avoiding any embedded biases from foreign training data that might undervalue context-specific factors like market vulnerabilities in emerging economies. Without such oversight, reliance on provider-defined alignment risks importing exogenous values that undermine the independence of regulatory decision-making.

[a] See, for instance, Anthropic's announcement on its core views on AI Safety: Anthropic, Core Views on AI Safety: When, Why, What and How ⟨https://www.anthropic.com/news/core-views-on-ai-safety⟩

Another potential problem resulting from character alignment is that the extent of alignment influences task suitability, since different LLMs embody varying philosophies that render them more or less appropriate for distinct phases of order writing. Highly safety-aligned models may emphasise guardrails against bias and promote consistent, rights-protecting outputs, making them well-suited for tasks requiring fairness and uniformity. In contrast, less-aligned commercial models or certain open-source systems like those based on Llama will exhibit fewer "corrective biases", allowing for unconstrained exploration that can be advantageous in creative or adversarial tasks.

To illustrate this contrast, consider a simple example in drafting regulatory orders related to insider trading under SEBI regulations. For the task of generating a reasoned order that applies penalties consistently across diverse market participants, a highly-aligned model might be preferable: prompted to draft a section on proportionate sanctions, it could produce text that explicitly weighs mitigating factors (e.g., the entity's size and intent) while embedding safeguards against discriminatory language, ensuring alignment with non-arbitrariness principles. However, for a task involving the exploration of potential loopholes in a proposed regulatory framework such as modelling how actors might exploit ambiguities in disclosure rules, a less-aligned model could prove more effective. Unburdened by heavy safety constraints, it might generate unconventional scenarios, such as hypothetical evasion strategies involving offshore entities, thereby aiding regulators in stress-testing drafts before finalisation.

These are just two examples where over-alignment may stifle innovative analysis in exploratory stages, while under-alignment could introduce risks in final drafting where precision and equity are paramount.

4.2 Integrating LLMs in quasi-judicial settings

Any technological solution deployed to assist order writing in quasi-judicial processes must necessarily align with the applicable principles of administrative law. Given that these principles have evolved from common law, the Constitutional scheme, and numerous judicial decisions, an exhaustive list is challenging to compile. For our analysis, we identify five principles applicable to order writing in quasi-judicial processes, as follows:[28]

1. **Application of mind**: Orders must demonstrate an application of the adjudicator's

[28]For a detailed discussion on the importance of aligning quasi-judicial processes with the principles of administrative law, see (Aggarwal, Patel, and Singh [n 18])

mind to the facts of the matter and the applicable law.[29]

2. **Reasoned order**: Orders must state reasons for a certain decision. Reasoned orders demonstrate application of mind. Courts have strongly emphasised the importance of providing reasons for decisions in quasi-judicial orders.[30] An unreasoned order is open to challenge in a higher forum and may be overturned in appeal, increasing costs for all parties involved.
3. **Non-arbitrariness**: The decision set out in an order, as well the process that led up to it, must not be arbitrary.[31] Decisions that are arrived at capriciously, or without a rational basis that is grounded in applicable precedent or statute, are likely to be set aside in appeal as arbitrary.
4. **Rules against bias**: The rules against bias mandate that all adjudicatory actions be impartial and unbiased.[32] The specifics of what constitutes bias in an adjudicator vary from case to case, and indeed, completely eliminating all preconceived notions may be impossible. Instead, administrative law recognises a threshold of a "reasonable likelihood of bias".[33] This rule prohibits adjudicatory authorities from trying matters they cannot decide with a reasonable degree of objectivity.
5. **Transparency**: Transparency is essential to ensure that the other principles enumerated here are satisfactorily met. In proceedings, it facilitates meaningful review of decisions by appellate bodies and other interested parties.[34]

   collectively, we call these principles "Applicable Law".

Applicable Law ensures and enables accountability in quasi-judicial processes, holding the adjudicator accountable for the soundness of the decision-making process. However, cerrtain limitations inherent to how LLMs work have the potential to detrimentally affect the adjudicator's ability to satisfy Applicable Law, in the following ways:

1. **Non-application of mind**: LLMs are probabilistic machines. Their method of reasoning is algorithmic and statistical, and their outputs are generated by predicting the most probable sequence of words based on patterns they identify in their training or grounding data, not by a deliberative, thoughtful process of applying legal principles to

[29] *MP Industries Ltd v Union of India* AIR 1966 SC 671.
[30] *Maneka Gandhi v Union of India* AIR 1978 SC 597.
[31] Ibid.
[32] *A K Kraipak v Union of India* AIR 1970 SC 150.
[33] *Jiwan K Lohia v Durga Dutt Lohia* (1992) 1 SCC 56.
[34] *M/s Kranti Associates Private Limited and another v Sh Masood Ahmed Khan and others* 2011 (273) ELT 345.

specific facts. Consequently, the content they produce may be contextually irrelevant or factually inaccurate. Directly incorporating such LLM output into quasi-judicial orders without meticulous human review and exercise of judgement would therefore constitute a non-application of mind, as it bypasses the essential human cognitive process of evaluating, interpreting, and judging the specific facts of a case against established legal norms.

2. **Black-box problem**: The inner workings of LLMs are complex and opaque. Their decision-making processes cannot be observed or mapped out, and therefore cannot be explained. The inexplicable nature of LLM functionality raises issues for the auditability of reasoning processes, directly impacting the transparency principle. Further, this very inexplicability presents a significant challenge to the reasoned-order requirement, as the basis for a certain output cannot be clearly articulated.
3. **Potential for bias**: Scholarly literature consistently establishes that LLMs risk replicating biases present in their training data, design, and use.[35] The rules against bias prohibit adjudicators from presiding over matters where there is a "reasonable likelihood" of bias. By virtue of its tendency to be biased, LLM output accepted without human review and application of mind may lead to unfairly discriminatory and arbitrary outcomes for parties, violating the rules against bias.
4. **Confabulation problem**: We noted that LLMs tend to generate fabricated, inaccurate, or contextually irrelevant results because of their predictive nature.[36] In legal contexts, there have been recorded instances of LLMs fabricating case law.[37] The output generated by an LLM is not arrived at based on objective, legally-sound criteria for decision-making. Should these materials be accepted without exercise of judgement or verification, the administrative law principle of non-arbitrariness could be undermined.
5. **Metacognition problem and the Dunning-Kruger effect in LLMs:** LLMs lack in metacognition (i.e., the ability to understand and acknowledge the limits of one's knowledge). Unlike a human adjudicator who can gather more information on a point of law or fact that is unfamiliar to them, LLMs prompted with questions beyond their knowledge base cannot similarly seek additional information. Singh and others (2024) point out that despite having inadequate information, LLMs tend to generate output, exhibiting a confidence-competence gap (commonly known as the Dunning-Kruger

[35]Isabel O Gallegos and others, "Bias and Fairness in Large Language Models: A Survey" [2024] arXiv ⟨https://arxiv.org/abs/2309.00770⟩.
[36]Banerjee (n 27).
[37]Singh (n 20).

effect in psychology).[38] Relying on output generated without adequate information therefore risks violating the application of mind principle and may lead to arbitrary decisions.

Additionally, we identify two logistical problems that must be addressed when using LLMs to assist in order writing processes:

6. **Training corpus**: To ensure the accuracy and relevance of output that the LLM-based application generates, it must be equipped with a carefully curated, sector-specific dataset of relevant laws, regulations, amendments, and precedent. However, these documents are often scattered around time and across jurisdictions, making them hard to collate and link. Further, some scholars argue that adjudicatory orders, like many other Indian legal documents can be long, noisy, and unstructured, making them poor training material.[39]
7. **Data privacy and security**: Quasi-judicial orders may sometimes contain sensitive data (such as personal data and intellectual property). Typically, breaches of this manner of sensitive data are likely to result in violations of intellectual property rights, information technology, and data protection laws. Adjudicatory officers and law clerks and other judicial research assistants must ideally maintain strict cybersecurity hygiene to avoid such implications. However, the risk of data breaches is heightened with the use of LLMs. Most commercially available pre-trained LLMs necessitate sending data to a remote server for output generation, thereby raising data security concerns for interested parties.[40] Considering this, there is a pressing need to address the problem of data privacy and security when using LLMs in order writing assistance.

In the next section, we refer to prior work on the requirements of good order writing, and try to divide the order writing process into those parts that permit the use of technology and those that may not. We then examine how the use of LLMs in quasi-judicial or judicial settings has been regulated in other jurisdictions. Then, we return to the problems of administrative law raised by any proposed use of LLMs in quasi-judicial settings, and propose a 'Problem-Solution-Evaluation' framework that could be applied to help ensure these problems and the requirements of good order writing are both addressed.

[38] Aniket Kumar Singh and others, "Do Large Language Models Show Human-like Biases? Exploring Confidence—Competence Gap in AI" (2024) 15(2) Information ⟨https://www.mdpi.com/2078-2489/15/2/92⟩.

[39] Abhinav Joshi and others, "IL-TUR: Benchmark for Indian Legal Text Understanding and Reasoning" [2024] arXiv ⟨https://arxiv.org/html/2407.05399v1⟩.

[40] Chun Hsien-Lin and Pu-Jen Cheng, "Assisting Drafting of Chinese Legal Documents Using Fine-Tuned Pre-trained Large Language Models" [2025] The Review of Socionetwork Strategies.

5 LLMs for good order writing

We now examine what requirements regulatory orders should satisfy. We then analyse the order writing process and evaluate whether LLMs may be used to assist human decision-makers in any parts of that process.

5.1 What makes for a good regulatory order?

Regulatory orders must satisfy several requirements to comply with applicable law and to withstand challenge in appeal. These requirements largely arise from two sets of sources:

- Principles of administrative law and natural justice, which extend generally to all regulatory actions, and
- Substantive legal and regulatory provisions, which vary by the law or regulation being applied in each situation.

Aggarwal, Patel, and Singh (2025) suggest that not satisfying these requirements raises concerns of legitimacy, efficiency, stakeholder interests, and legal requirements in relation to regulatory order writing.[41] They posit that these can be addressed through the following principles. Orders must:

- clarify the basis for the exercise of regulatory power,
- clearly identify the legal provisions that are purportedly violated,
- provide detailed background and contextual information,
- provide well-reasoned justifications for findings of fact and law,
- specify the regulatory action being taken and correlate it with the findings of fact and law, and
- describe the rights of review or appeal of the persons involved

These principles can be addressed by fulfilling the following four sets of requirements:

- Information Requirements
- Structural Requirements
- Substantive Requirements
- Stylistic Requirements

Fulfilling these requirements is a non-trivial task, for several reasons:

[41] Aggarwal, Patel, and Singh (n 18).

- There is a lack of standardised order writing practices and materials in Indian regulatory institutions. There is no specific articulation of order writing requirements in Indian law applicable to all regulators, beyond the general rule in Section 4(1)(d) of the Right to Information Act, 2005 which requires that every public authority "provide reasons for its administrative or quasi-judicial decisions to affected persons."[42] These exacerbate the problem.
- Order writing practices vary by regulator and within regulatory institutions. There is a lack of consistency in the kinds of information included in regulatory orders, as well as the manner in which orders are structured.
- The high frequency of changes to regulatory instruments and applicable law creates challenges to achieving accuracy in the substantive legal aspects of orders.
- There is also a lack of clarity on whether the principles of precedent and the requirement of citing previous decisions applies to regulators under Indian law, and if so, in what manner.
- Finally, the stylistic requirements of clarity and simplicity are often not met, and the kind of language employed varies based on the author of each order.

We have not noticed any publicly available materials that provide clarity on these issues, nor any materials that provide guidance on how regulatory orders and order writing practices may be standardised. We have also not noticed any publicly available material that indicates how regulatory officers are trained specifically on good order writing, and not just the substantive and procedural aspects of their institution's regulatory domain.

While providing such guidance, training, and materials could help address the problem, we suggest that there is also a need to explore how technological tools could be used to aid order writers in their task.

5.2 LLMs in the order writing process

In order to evaluate whether technological tools can be used to assist order-writers, it is important to acknowledge that good order writing by regulators is not a monolithic activity. It requires careful work over several steps. These steps include:

- **Documentation and referencing**: Ensuring detailed records of proceedings are considered and referred to,
- **Research**: Involves precedent checking, study of applicable law, etc.,

[42] Section 4(1)(d), Right to Information Act 2005.

- **Decision-making**: Application of law to facts, arriving at conclusions and determining consequences such as sanctions,
- **Structuring**: Writing a logically well-structured document that strengthens the force of analysis and argument,
- **Reviewing**: Ensuring draft orders are complete in all respects, and that they present strong logical arguments, and
- **Clear writing**: Ensuring an order is written in a comprehensible and executable manner.

Some of these steps may not be amenable to technological involvement at all, while others may permit the use of technological aids to varying degrees. As an example, it may be more straightforward to argue that technological aids can be developed to assist research. On the other hand, the administrative law rule requiring application of mind by the decision-maker may prohibit abdicating to technology the responsibility of taking decisions that affect the rights of involved parties.

The use of GenAI or LLMs is not the only manner in which technology could augment or improve the order writing process. Some steps involved in good order writing that do permit the use of technological aids may not require the use of such types of technology. They may instead benefit from the use of of rules-based systems, or other systems that are more reliable and predictable, and which do not generate probabilistic outcomes. These tools could include:

- conditional or logical forms that employ decision trees to guide the author's decision-making through a series of steps customised to the nature of the regulatory matter being considered, and
- templates that require the order writer to draft orders in a specified sequence, to help ensure that all necessary elements are present and in the correct sequence.

We suggest that the use of LLMs to aid regulatory order writing would be helpful and could complement such basic technologies. Some steps that are part of the overall order writing exercise may benefit from the use of LLM-based technological aids. As examples, LLMs could, with the necessary guardrails in place:

- be used to review draft orders for completeness, and could identify if any essential information that should be included in the order has been omitted,
- check draft orders to ensure they are structured in a standardised and logical manner.

LLMs might offer the author suggestions for restructuring the draft if necessary.

- identify, extract, summarise, and organise the most relevant data from pleadings and legal databases. LLMs could be used as a supporting/ verification mechanism to ensure this has been done in the correct manner,
- be asked to assist with the process of identifying applicable precedent from a precedent bank, based on whether the facts under consideration are analogous to the material facts in previous orders, and
- ensure the stylistic requirements of good order writing are met, and make suggestions for changing complex or difficult to read language in the draft, so as to ensure readability and clarity.

The use of LLMs in such applications has associated risks. We therefore study how they have been used in other jurisdictions for similar tasks, and what principles or rules have been developed to constrain the potential harm that such use of LLMs in particular, and AI in general may cause.

6 International practices and design principles

Some jurisdictions have considered how AI can be used in various aspects of governance by the state, and have articulated frameworks to govern such use. These frameworks offer insight on the possibility of integrating AI systems into quasi-judicial functions in a manner that complies with Applicable Law. These frameworks are not necessarily specific to LLMs; they discuss how AI can be used for different government functions, such as surveillance, profiling, regulatory enforcement, and adjudication. For example, in December 2024, the Administrative Conference of the United States published a report and adopted a recommendation on the use of algorithmic tools in regulatory enforcement.[43] As another example, the Higher Council of the Judiciary in Colombia has adopted an agreement for "the use and implementation of generative AI in the Judicial Branch."[44] These frameworks therefore offer insight on how guardrails can be designed for the use of AI in government functions in general, including adjudication.

[43] Administrative Conference of the United States, *Artificial Intelligence and Regulatory Enforcement* (Final Report, Administrative Conference of the United States) ⟨https://www.acus.gov/sites/default/files/documents/AI-Reg-Enforcement-Final-Report-2024.12.09.pdf⟩; Administrative Conference of the United States, *Using Algorithmic Tools in Regulatory Enforcement* (Recommendation 2024-5, Recommendation 2024-5, Administrative Conference of the United States 2024) ⟨https://www.acus.gov/document/using-algorithmic-tools-regulatory-enforcement⟩; David Freeman Engstrom and others, *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (Report, Administrative Conference of the United States 2020) ⟨https://law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf⟩.

[44] Higher Council of the Judiciary of Colombia, Agreement PCSJA24-12243 (December 2024).

We have studied such frameworks from Australia, Canada, China, Colombia, the European Union, India, the United Kingdom, and the United States. The frameworks are listed and briefly described in **Annexure A**. These frameworks also articulate certain principles, such as human oversight, that circumscribe the design and use of AI in quasi-judicial contexts. We identify and extract some common principles present across these frameworks, and describe them below:

- **Non-discrimination and fairness:** An AI tool should not develop or intensify discrimination and bias or impede the fairness of an adjudicatory process or its outcome.
- **Transparency and explainability:** An AI tool should allow a decision-maker to understand how the system works and its reasoning, and evaluate its functionality.
- **Human oversight:** An AI tool should be designed as a tool for assistance with, not replacement of, judicial functions. A decision-maker should continue to be responsible for a decision, and must review and analyse any AI-generated output and information.
- **Security and confidentiality:** An AI system should have limited access to sensitive or confidential information, and there must be safeguards to ensure the protection of the right to privacy and confidentiality.
- **Respect for fundamental rights:** The use of AI by an adjudicatory authority must respect, protect, and promote fundamental rights.
- **Risk management:** AI systems should be designed with appropriate and targeted risk management measures designed to identify, assess, and mitigate potential risks, such as conflabulations, bias, or violations of fundamental rights.

  collectively, we identify these common principles for the use of AI in quasi-judicial settings as the "Design Principles".

The Design Principles demonstrate how these jurisdictions have sought to mitigate the inherent limitations of LLMs and develop tools that comply with applicable law, including administrative law.

Table 2 identifies the Design Principles included in each jurisdiction's framework(s).

Table 2 Design principles and corresponding Applicable Law

This table illustrates the inclusion of the Design Principles in other jurisdictions' frameworks and Applicable Law that relate to each Design Principle.

Design principle	**Australia***	**Canada***	**China***	**Colombia***	**European Ethical Charter and EU AI Act***	**Germany**	**India†**	**UK***	**USA***	**Applicable Law**
Non-discrimination and fairness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Non-arbitrariness; Rules against bias
Transparency and explainability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Unclear	Application of mind; Reasoned order; Transparency
Human oversight	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Yes	Application of mind
Security and confidentiality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-
Respect for fundamental rights	Yes	Yes	Unclear	Yes	Yes	Unclear	Yes	Unclear	Unclear	Non-arbitrariness; Rules against bias
Risk management	Unclear	Yes	Yes	Yes	Yes	Unclear	Yes	Unclear	Yes	-

* Frameworks address use of AI by judicial authorities. USA's executive memorandum addresses use of AI by federal agencies.

† Based on the Supreme Court's White Paper on Artificial Intelligence and Judiciary (November 2025), the Ministry of Electronics and Information Technology's Governance Guidelines (November 2025), and a policy issued by the High Court of Kerala for the use of AI tools by the district judiciary (July 2025).

Note: Policy frameworks for Argentina, Brazil and Estonia not available in English.

Table 2 also demonstrates that the incorporation of these Design Principles can help in creating solutions that satisfy the requirements of Applicable Law.

Building on these Design Principles, several jurisdictions have also integrated AI into their judicial and/or quasi-judicial functions. For example, in the United States, the Patent and Trademark Office uses AI tools to review patent and trademark applications.[45] **Annexure B** lists other examples that demonstrate the integration of AI in judicial and quasi-judicial functions. Both, the Design Principles and the exemplar uses of AI in other jurisdictions inform our understanding of how AI systems can be designed and evaluated in compliance with Applicable Law.

7 LLMs for order writing: Problems, solutions, and evaluations

We now describe a three-part conceptual framework for the use of LLMs in assisting order writing processes. In Section 4, we identified a set of problems that need to be addressed when using LLMs in such situations. In this Section, we propose potential solutions to address each of these problems. We then suggest an additional layer of evaluations designed to assess the efficacy of the solutions and facilitate iterative refinement. Together, these constitute what we call the 'Problem-Solution-Evaluation' framework.

Tables 3 and 4 illustrate our conception of the PSE framework. Table 3 identifies Applicable Law potentially impacted by each identified problem. It then demonstrates how the PSE framework can be operationalised to reduce detrimental effects on Applicable Law. Table 4 shows how the PSE framework can be used to address the logistical problems with using LLMs for order writing assistance.

[45] Engstrom and others (n 43).

Table 3 Applying the Problem-Solution-Evaluation framework

This table illustrates how the PSE framework can be operationalised to align the design, development and use of LLMs for order writing assistance with the requirements of Applicable Law. Annexure C elaborates on the methods proposed in this table.

Problem	Applicable law	Solution	Evaluation
Non-application of mind	Non-application of mind; Failure to provide reasons Arbitrariness	Interface Checkpoints Confidence Score Display Dual-Prompt Pipelines Functionality Limitation Constraint Enforcement Workflow Design for Review Role-Based Access	Edit Rate Turnaround Time (TAT) Prompt Divergence Rate Coherence Score
Black-box problem	Failure to provide reasons; Transparency	Chain-of-thought prompting Input Token Influence Identification Symbolic Reasoning Systems Traceability tools Visualisation Simplified model explanations	Clarity rating Audit Trail Incidence Document Traceability Rate
Potential for bias	Rules against bias Arbitrariness	Data Preprocessing Bias penalisation Domain-specific content filters Automated Bias Flagging Tools Establishment of Legal Fairness Criteria Mandatory Periodic Benchmarking	Bias Flag Rate Override Percentage Fairness Benchmark Scores
Confabulation problem	Non-application of mind Failure to provide reasons Arbitrariness	Retrieval Augmented Generation Post-Generation Verification Legal Knowledge Graph Integration Mandatory reviewer verification Watermarking for traceability Communicate technical limitations	Secondary LLM “Judge” for Fact-Checking End-to-End Evaluation Tools Hallucination Rate Retrieval Precision@k/ MRR NLI Coherence Checks Self-Consistency Rate
Lack of metacognition	Non-application of mind Arbitrariness	Prompt engineering LLM as a judge Iterative improvement from feedback	Closeness Metric Human evaluation on overconfidence in output

Table 4 Applying the Problem-Solution-Evaluation framework (Continued)

This table illustrates how the PSE framework can be operationalised to solve for logistical problems involved in using LLMs for order writing assistance. Annexure C elaborates on the methods proposed in this table.

Problem	Applicable law	Solution	Evaluation
Training corpus	NA	Adaptive Scraping Frameworks Sector-specific pre-training Structured Entity Extraction and Legal Knowledge Graphs Isolated Model Containers Source inclusion Perplexity tracking Legal Retrieval Benchmarking Curate sector-specific legal databases	Crawl coverage OCR Error Reduction Validation perplexity Retrieval lift
Data security and privacy	NA	Stringent access control Synthetic supervision-based PII detectors NLP filters for information masking Isolated Model Containers On-premise infrastructure	Unauthorised access attempts Mean Time To Remediation (MTTR) Penetration Test Pass Rate PII Detection Accuracy

7.1 Solutions

Having outlined the potential detrimental impact that the inherent limitations of LLMs can have in light of Applicable Law, we now propose a framework of solutions to minimise it. To re-emphasise, solutions are so called because they are mechanisms and processes that help identify the site of problems and facilitate their management. They are not, in our usage, intended as cure-alls that remove incidences of the problem entirely.

The solutions framework must focus on all levels of the system's operation: interventions within the system's core, the interactions between the user and the system, as well as broader, ecosystem-level measures that influence the system, the user, and the outcomes they collectively generate. Accordingly, we propose that the problems we identified in Section 3 can have solutions that fall into one of three categories, namely:

1. **Technical solutions**: Ways in which the workflow of the LLM-based application can be designed.
2. **Design solutions**: Ways in which the application's interface with the user must be designed to optimise human control.
3. **Systemic solutions**: Measures that need to be taken at the organisational and sectoral level by regulatory and adjudicatory bodies that use such applications, and by other stakeholders, collectively.

7.1.1 Technical solutions

The workflow and functioning of an LLM-based application intended to assist in order writing must be carefully designed to address the problems we have outlined above. To do this, the first layer of solutions we propose comprises what we refer to as 'technical solutions'. We define technical solutions as involving two kinds of measures: the deliberate design of the LLM's inner workflow, and its specific computational techniques.

For instance, consider the confabulation problem in an LLM-based application used to identify relevant precedent. Hannah and others (2024) propose that since legal research tasks such as this are "knowledge-intensive", LLMs need to be supplemented with the necessary grounding information to improve their accuracy and relevance to context.[46] They propose that this grounding information can come from a Legal Knowledge Graph, integrated into the application's workflow to facilitate information retrieval for the LLM to summarise or

[46] George Hannah and others, "A Prompt Engineering Approach and a Knowledge Graph based Framework for Tackling Legal Implications of Large Language Model Answers" [2024] arXiv ⟨https://arxiv.org/pdf/2410.15064⟩.

display, with appropriate citations to the source document. This is an example of a technical solution.

7.1.2 Design solutions

The next layer of solutions should focus on the user's interaction with the application. We argue that an order writing assistance application must prioritise user control, mandating human verification and approval for every step of the LLM's functioning and output generation. This approach aligns with international best practices in design principles that strongly emphasise user control.[47]

A user-in-control design places a check on both the application and its user. For instance, an adjudicator who seeks to include the most relevant precedent in an order, must retain final authority over which precedential positions are prioritised over others. The application must, by design, require the adjudicator to manually cross-verify and sign off on each identified case before it is included in the final order. When accepting or deviating from the LLM's suggestions for applicable precedent, the adjudicator must be compelled by design, to provide justifications for their treatment of a certain output. This integration of review into the interface ensures that the adjudicator is held accountable to apply their mind, while also ensuring the verification of output before it is accepted.

7.1.3 Systemic solutions

The third layer of solutions comprises broader, ecosystem-level measures governing the LLM application's development, deployment, and use. These contextual measures must involve and account for the participation of the stakeholders in the order writing process.

For instance, stakeholder participation and collaboration is essential to building an LLM-based application used to assist order writing processes. The resources necessary to develop a useful and efficient LLM-based legal application (such as a training dataset of applicable laws, regulations and precedent) are scattered across time periods and jurisdictions.[48] These need to be aggregated through collaborative effort for the development of the application. This is an example of a systemic solution.

[47] See, Section 4 above.
[48] Hsien-Lin and Cheng (n 40).

7.2 Evaluations

We propose that the solutions framework must be supplemented by a continuous evaluation framework. These evaluations will provide technical, design, and systemic checks, assessing the solutions' efficacy in aligning the LLM-based application's functioning and its output's use with the requirements of Applicable Law. This also gives stakeholders the chance to enhance the application's robustness and utility while simultaneously refining their interactions with it to improve outcomes.

We offer four possible architectural frameworks to design such an evaluation mechanism:

1. End-to-end evaluation
2. Component-wise evaluation
3. Human-in-the-loop evaluation
4. Automated evaluation frameworks

7.2.1 End-to-end

An end-to-end evaluation framework focuses on the overall quality of the application's functioning, by examining the final output. The purpose of this evaluation strategy is to measure the application's performance against high-level criteria, such as factual accuracy, coherence, completeness, and adherence to relevant standards of research and documentation. A systemic pre-cursor to this manner of evaluation would be to adopt a common set of standards for well-researched and written orders, based on the principles of administrative law, best practices and the specificities of each adjudicatory body. The benchmarking framework for a well-written regulatory order proposed by Aggarwal and others (2025) is one such example of this.[49] Further, frameworks like DeepEval's *G-Eval* allow users to evaluate LLM-output using "*any custom criteria*".[50]

7.2.2 Component-wise

A component-wise evaluation framework involves dissecting the system into its constituent parts and evaluating each component independently. This approach allows for the identification of specific bottlenecks or areas where improvements are needed within the overall system. This allows both users and developers the opportunity to identify specific pain points

[49] Aggarwal, Patel, and Singh (n 18).

[50] G-Eval (DeepEval,) ⟨https://deepeval.com/docs/metrics-llm-evals⟩; Yang Liu and others, "G-EVAL: NLGEvaluation using GPT-4 with Better Human Alignment" [2023] arXiv ⟨https://arxiv.org/pdf/2303.16634⟩

in the functioning of the system, and ways to work around them. This form of evaluation, too, requires an established set of standards to be agreed upon and adopted. Benchmarking systems, such as Joshi and others (2024)'s IL-TUR,[51] can serve as starting points for this. Such benchmarking systems need to be tailored to suit the requirements of each order writing workflow.[52]

7.2.3 Human-in-the-loop

This evaluation method integrates human reviewers' expertise (similar to Reinforcement Learning from Human Feedback) to capture qualitative feedback.[53] This feedback covers subjective aspects such as reasoning, logical coherence, and a nuanced understanding of administrative law requirements in order writing. This approach offers invaluable insights that automated metrics might not capture.

7.2.4 Automated evaluation

Finally, one layer of evaluations can be outsourced to LLMs or other AI models acting as judges. These models would automatically score and assess generated content based on predefined criteria, offering benefits of scalability and efficiency in evaluating large volumes of documents.[54] However, the reliability of such frameworks depends greatly on carefully designed evaluation prompts and validating the judge's performance through human verification to mitigate potential algorithmic bias.

Put together, the PSE framework helps:

- identify misalignments between the requirements of administrative law and the constraints of technological solutions used to assist in order writing processes;
- devise a broad range of solutions and approaches to meaningfully address and reconcile the problems that arise with the use of technologies like LLMs in order writing; and
- continuously evaluate the efficacy of the solutions, and iteratively make improvements to the design of the application, to align it with the requirements of administrative law.

[51] Joshi and others (n 39).

[52] See, G-Eval (n 50); Liu and others (n 50)

[53] See, Long Ouyang and others, "Training language models to follow instructions with human feedback" [2023] arXiv ⟨https://arxiv.org/pdf/2203.02155⟩

[54] Aman Madaan and others, "Self-Refine: Iterative Refinement with Self-Feedback" [2023] arXiv; Yuntao Bai and others, "Constitutional AI: Harmlessness from AI Feedback" [2022] arXiv

7.3 The PSE framework in action: Solving and evaluating the black-box problem

Having described each component of the framework, we will now illustrate how we envision these components working together. Consider, for instance, the black-box problem.[55] The opacity of reasoning paths in LLMs affects the auditability of LLM systems. Specifically, when the logical steps or data points leading to a particular output remain opaque, discerning the underlying rationale becomes challenging. This directly challenges the administrative law requirements for transparency and reasoned orders. To improve the auditability of LLMs used for order writing assistance, specific technical, design, and systemic solutions can be implemented as follows:

Technical solutions

- *Chain-of-thought prompting*: Apply methodologies such as chain-of-thought prompting to compel models to explicitly articulate their intermediate reasoning steps.
- *Input Token Influence Identification*: Utilise interpretability techniques[56] like *LIME* (Local Interpretable Model-agnostic Explanations)[57] and SHAP (SHapley Additive exPlanations)[58] to identify which input token influenced each segment of the generated output.
- *Symbolic Reasoning Systems*: Combine LLMs with symbolic reasoning engines that enforce rule-based logic.
- *Traceability tools*: Incorporate specialised tools that log and record the usage of source documents during the text generation process, to create an auditable trail of information provenance.

Design solutions

- *Visualisation*: Employ visualisation tools to represent attention layers and the relevance of source materials. This can reveal which parts of the input data the model focused on, thereby highlighting the evidential basis for its conclusions.

[55] The application of the framework to the other problems we identified in Section 2.2 is elaborated on in **Annexure C.**

[56] The term 'interpretability' refers to the direct understanding of the internal processes of an AI model. The term 'explainability', on the other hand, refers to providing post-hoc explanations for an AI model's output, in line with human reasoning, while its internal mechanisms remain opaque to humans. See, Ričards Marcinkevičs and Julia E Vogt, "Interpretable and explainable machine learning: A methods-centric overview with concrete examples" (2023) 13(3) WIREs ⟨https://doi.org/10.1002/widm.1493⟩

[57] marcotcr / lime ⟨https://github.com/marcotcr/lime⟩.

[58] shap ⟨https://github.com/shap/shap⟩.

Systemic solutions

- *Simplified model explanations*: Publish simplified explanations of model functionality to demystify their basic operations for a broader audience, including regulators, adjudicators, and other stakeholders.

Once various kinds of solutions are put in place, they must be supplemented with a layer of evaluations that test the efficacy of the solutions in meeting the requirements of Applicable Law affected by the black-box problem. Some metrics for this evaluation could include:

- *Clarity rating*: A regulator-assessed rating of the clarity and utility of the model's intermediate reasoning, particularly focusing on the interpretability and usefulness of chain-of-thought outputs.
- *Audit Trail Incidence*: Quantifies the number of inconsistencies or anomalies flagged during periodic audit reviews of the model's operations and its generated outputs. A lower incidence rate indicates higher transparency and auditability.
- *Document Traceability Rate*: Represents the proportion of generated content for which attributable source references can be accurately identified. A high traceability rate is indicative of robust factual provenance, improving transparency.

Insights from these evaluations must then be periodically integrated into the solutions layer to better align the application's functionality with Applicable Law, improving its auditability. This iterative process allows the PSE framework to be operationalised, effectively addressing problems arising from LLM use in order writing assistance, such as those identified in Section 2.2.

8 Recommendations

Any use of LLMs in quasi-judicial settings should be based on a robust Problem-Solution-Evaluation framework. However, this framework should be considered alongside the question of which specific steps of the order writing process are amenable to the use of technology. Considering both these, as well as the principles of Indian administrative law, we make some recommendations about how LLMs can be used in quasi-judicial settings. Our recommendations are divided into two broad phases: the first is the 'Build' phase, in which we make recommendations relating to how systems are designed and developed, and the downstream effects of such design and development; the second is the 'Use' phase, in which we make recommendations about how such systems are used, and the interactions between humans

and such systems.

'Build' phase recommendations:

While designing and developing systems that leverage LLMs as tools to aid order writing, developers should ensure that:

- the Problem-Solution-Evaluation framework is a core component of the system's design,
- the system offers suggestions or options to the user, and not definitive 'answers' that comprise decisions on substantive matters,
- systems are used to 'review' draft orders, and not to create those drafts in the first instance,
- thorough documentation is prepared and made publicly available, describing, at the least, the training corpus used, the algorithms used to train the system, and any assumptions the developers have made about the processes the system aims to assist with,
- the system is used to 'shadow' human decision-makers for a substantial period, its performance compared to that of humans, and any remedial measures necessary are taken prior to the deployment of the system in a 'live' environment, and
- the system design ensures fact-checking and the citation of sources for all suggestions made to the user.

'Use' phase recommendations:

The use of any system that leverages LLMs in quasi-judicial processes must be complemented with:

- the use of means that ensure and demonstrate application of mind by the human decision-maker; one example of this could be the requirement for decision-makers to record internal notes explaining why they agree or disagree with the system's suggestions,
- 'on-premise' deployment or other appropriate measures that help ensure data security,
- training and periodic re-training for adjudicators on the uses and potential dangers of systems that leverage AI in general and LLMs in particular, and
- periodic system audits and checks, and the use and frequent upgradation of benchmarking and quality assurance measures.

9 Conclusion

Technology can be used to augment regulatory state capacity in several aspects of the order writing process. In order to understand what aspects these are, and how technology may be used in them, it is important to deconstruct the order writing process into its constituent activities, and identify the requirements that the principles of administrative law impose on each of these activities. This permits the designing of a considered approach that accounts for both, the limitations and problems inherent in the form of technology under consideration, and the requirements of Applicable Law.

We suggest that a considered approach of this nature can help balance extreme suggestions about the use of AI and LLMs in quasi-judicial processes. Substituting the human decision-maker with an automated tool may violate Applicable Law and may also be undesirable for other reasons; we do not suggest such a substitution. We propose the approach described in this paper, and the PSE framework in particular, as means to augment human decision-makers' capabilities and ensure the better realisation of the rule of law in Indian regulatory quasi-judicial processes. It is important to emphasise that current problems in regulatory order writing do not seem to arise from lack of capacity, in the sense that there is a shortage of personnel at regulatory institutions. This may be true in some instances, but the larger problem would seem to be the lack of availability of tools and resources for adjudicators. This problem cannot be solved by training alone; unless a decision-maker has all the tools necessary to make an informed, well-reasoned decision, it is difficult to expect such decisions. Our suggestions relate to how technology can be used to fill this gap.

Our recommendations are based on our study and understanding of Applicable Law and the current limitations of GenAI and LLMs. We propose to conduct further research to test their validity. This could include setting up prototype systems that foreshadow what an eventual order writing-assistance system might look like, and comparing their suggestions and performance against current practices. Any eventual rollout of such a system would also have to include a training programme for regulatory officials on the safe and efficacious use of such technologies.

Finally, we are cognisant of the fact that our recommendations are likely to be impacted by the ever-evolving landscape of technological progress. As GenAI and LLMs evolve, we will have to continuously re-evaluate the PSE framework, and our recommendations.

A Annexure A

We focus on the use of AI in judicial and quasi-judicial functions, and study the following frameworks:

- **Australia**: In November 2024, a report by the Senate Select Committee on Adopting Artificial Intelligence described the role of automated decision-making.[59] The Commonwealth Ombudsman issued a practice guide in March 2025 to ensure compliance with administrative law principles and privacy requirements in the use of automated systems.[60]
- **Canada:** The Canadian Government has issued a directive on automated decision-making systems, which is intended to ensure that such systems comply with core principles of administrative law.[61]
- **China**: In December 2022, the Supreme People's Court of China issued the Opinion on Regulating and Strengthening the Applications of Artificial Intelligence in the Judicial Fields (Opinion) for "integration of artificial intelligence with judicial work, [and to] deepen the construction of smart courts, and strive to achieve a higher level of digital justice."[62] The Opinion has an expansive list of instances in which AI may be used, such as classification and categorisation of cases, case information crawling, review of evidence, and smart push of relevant laws, regulations, and precedent, issuing warnings for deviation from adjudicative norms, reviewing procedurally terminated enforcement cases, inspecting judicial irregularities, and identifying risks of judicial corruption.[63] AI may also be used to generate and review legal documents, recommend judicial resolutions, and predict litigation outcomes.[64]
- **Colombia**: The Higher Council of the Judiciary in Colombia adopted Agreement

[59]Senate Select Committee on Adopting Artificial Intelligence, Final Report: Adopting Artificial Intelligence (ISBN 978-1-76093-759-1; Released 26 November 2024, November 2024) ⟨https://parlinfo.aph.gov.au/parlInfo/download/committees/reportsen/RB000470/toc_pdf/SelectCommitteeonAdoptingArtificialIntelligence(AI).pdf⟩.

[60]Commonwealth Ombudsman and Australian Information Commissioner, Automated Decision-Making – Better Practice Guide (Published by the Office of the Commonwealth Ombudsman (Australia),) ⟨https://www.ombudsman.gov.au/__data/assets/pdf_file/0025/317437/Automated-Decision-Making-Better-Practice-Guide-March-2025.pdf⟩.

[61]Treasury Board of Canada Secretariat, Directive on Automated Decision-Making (Active since April 1, 2019; applies to systems developed or procured after April 1, 2020, 2024) ⟨https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592⟩.

[62]Supreme People's Court of the People's Republic of China, The Supreme People's Court: The Opinions on Regulating and Strengthening the Applications of Artificial Intelligence in the Judicial Field (2022) (English translation available at China Justice Observer, December 2022).

[63]Ibid.

[64]Ibid.

PCSJA24-12243 in December 2024 (Agreement) for "the use and implementation of generative AI in the Judicial Branch."[65] The Agreement specifies how AI may be used in administrative management or to support judicial management (for example, assisted drafting of administrative texts such as letters and reports). AI may also be used to: (1) reference precedent, (2) support the thematic classification of documents, (3) review the completeness of documents, (4) assist in drafting procedural orders, (5) assist in the improvement of order drafting, (6) simulate "case-specific decision scenarios", (7) assist with "tasks that have an impact on the work of motivating judicial decisions", (8) assist with summaries of facts and testimony, and (9) analyse decisions in proceedings with "standardised and recurring legal problems".[66] Some of these uses are subject to higher standards of transparency and accountability than others.

- **European Union:** The Artificial Intelligence Act (AI Act) was adopted in June 2024. It regulates the use of AI in the European Union through a risk-based classification, in which the use of AI by the judiciary is considered 'high-risk.'[67] The European Commission for the Efficiency of Justice of the Council of Europe issued the "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment" in December 2018.[68]
- **India:** In November 2025, the Supreme Court of India's Centre for Research and Planning issued a White Paper on Artificial Intelligence and Judiciary. This White Paper emphasises core principles, such as human oversight and confidentiality, and outlines several potential uses of AI by the judiciary, including for computing "financial components in complex commercial disputes" and generating templates for "documents, speeches, presentations, and similar materials".[69] In March 2025, the Ministry of Law and Justice issued a press release on the use of AI by the judiciary and law enforce-

[65] Agreement PCSJA24-12243 (n 44). This agreement was adopted pursuant to a decision by the Constitutional Court in August 2024, which involved a judge who used generative AI to write a decision on health insurance. The primary question was whether such use of AI violated the fundamental right to due process of law. The Court held that it did not, because the AI did not replace judicial reasoning and was only used to transcribe the decision after the judge had arrived at a conclusion. See Institute for the Development of Society (IDS), In a New Ruling, Colombia's Constitutional Court Analyzes the Use of AI Tools by the Country's Judiciary (August 2024).

[66] Agreement PCSJA24-12243 (n 44).

[67] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on harmonised rules on artificial intelligence (Artificial Intelligence Act) (Classifies AI used in judicial decision-making as high-risk, 2024).

[68] European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment (2018).

[69] Supreme Court of India, *White Paper on the Use of Artificial Intelligence in the Indian Judicial System* (White Paper, Supreme Court of India 2025) ⟨https://cdn.s3waas.gov.in/s3ec0490f1f4972d133619a60c30f3559e/uploads/2025/11/2025112244.pdf⟩.

ment, particularly in the context of the e-Courts Project.[70] Some of the potential uses of AI include identifying precedent and offering predictive analyses, which can assist "judicial officers to formulate more informed decisions and develop effective case strategies."[71] In July 2025, the Kerala High Court adopted a policy for the use of AI tools by the district judiciary, which clarifies that "AI tools shall not be used to arrive at any findings, reliefs, order or judgment under any circumstances".[72]

- **UK**: The Courts and Tribunals Judiciary has issued guidance on the use of AI by judicial officers, noting that AI can be used for summarising text, writing presentations, and administrative tasks.[73] The guidance recommends not using AI for legal research, unless the results are independently verifiable, or for legal analysis.[74]
- **USA**: Several executive orders over the last few years have outlined the use of AI by federal agencies. Most recently, a memorandum by the Office of Management and Budget classified the use of AI for adjudication by regulatory agencies as "high-impact."[75] Additionally, the Artificial Intelligence Rapid Response Team, comprising justices, administrative judges, and administrative officers, developed the Guidance for Use of AI and Generative AI in Courts in August 2024.[76] In October 2024, the Delaware Supreme Court adopted an Interim Policy on the Use of GenAI by Judicial Officers and Court Personnel.[77]

The relevant frameworks from Argentina, Brazil, Estonia, and France are not available in English and so we have not been able to analyse them.

[70] Ministry of Law and Justice, "Digital Transformation of Justice: Integrating AI in India's Judiciary and Law Enforcement" (n 17).

[71] Ibid.

[72] High Court of Kerala, *Policy Regarding the Use of Artificial Intelligence Tools in the District Judiciary* (Official Memorandum, HCKL/7490/2025-DI-3-HC KERALA,, High Court of Kerala 2025) ⟨https://images.assettype.com/theleaflet/2025-07-22/mt4bw6n7/Kerala_HC_AI_Guidelines.pdf⟩.

[73] Courts and Tribunals Judiciary, *Artificial Intelligence (AI) Guidance for Judicial Office Holders* (Guidance, Courts and Tribunals Judiciary 2025) ⟨https://www.judiciary.uk/wp-content/uploads/2025/10/Artificial-Intelligence-AI-Guidance-for-Judicial-Office-Holders-2.pdf⟩; Courts and Tribunals Judiciary, *Artificial Intelligence (AI) Guidance for Judicial Office Holders* (Guidance, Courts and Tribunals Judiciary 2025) ⟨https://www.judiciary.uk/wp-content/uploads/2025/04/Refreshed-AI-Guidance-published-version-website-version.pdf⟩.

[74] Courts and Tribunals Judiciary, *Artificial Intelligence (AI) Guidance for Judicial Office Holders* (n 73); Courts and Tribunals Judiciary, *Artificial Intelligence (AI) Guidance for Judicial Office Holders* (n 73).

[75] Office of Management and Budget, Executive Office of the President, US, M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust ⟨https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf⟩.

[76] Conference of Chief Justices National Center for State Courts and Conference of State Court Administrators, Guidance for Use of AI and Generative AI in Courts (2024).

[77] Supreme Court of Delaware, Interim Policy on the Use of GenAI by Judicial Officers and Court Personnel (2024).

B Annexure B

Some examples of the use of AI by judicial and quasi-judicial authorities are below:

1. In Argentina:
    - The Court of Justice of the Province of San Juan has approved a protocol for the use of generative AI for the judiciary.[78]
    - The Superior Court of Justice of the Province of San Luis has also approved the use of IURIX Mind and IURIX Cloud Native for judicial management.[79]
    - The Secretary of Jurisprudence for the Superior Court of Justice of La Pampa has developed 'Genaro', a tool to review rulings for length, complexity, style and clarity. Its responses rely on a drafting guide prepared by the Court.
    - Prometea, an AI tool, has been used by several courts (for example, the Court of Enforcement Proceedings, Bankruptcy and Insolvency in the Chaco province) and can search for precedent, classify documents, and predict case outcomes. Promotea does not replace the human decision-maker, and instead proposes outcomes based on previous decisions on analogous facts. Prometea has been used to pass judicial orders on housing rights, resolve cases involving driving under the influence, and sentencing in tax matters.[80]
2. In Brazil, the Judiciary has introduced the APOIA system, which integrates multiple AI models to assist courts with drafting reports, summarising case files, and identifying relevant legal provisions.[81] AI tools have been adopted by at least 47 courts,[82] including:
    - INACIA, a tool combining pre-trained LLMs and search engines, which is used by auditors in the Federal Court of Accounts to suggest "adjudication directions to court decision-makers – consisting of structured reasoning around cases' claims, evidence, and relationship to legal provisions in Brazilian law."[83] The implemen-

[78] Milagros Denise Tallarico, Argentina Approves Protocol for the Use of Generative Artificial Intelligence in the Judiciary (November 2024) ⟨https://www.theworldlawgroup.com/membership/news/news-argentina-approves-protocol-for-the-use-of-generative-artificial-intelligence-in-the-judiciary-1⟩.

[79] Redacción InnovaciónDigital360, "Poder Judicial argentino: por qué la IA puede ser clave para su funcionamiento" [2024] InnovaciónDigital360 ⟨https://www.innovaciondigital360.com/i-a/poder-judicial-argentino-por-que-la-ia-puede-ser-clave-para-su-funcionamiento/⟩.

[80] Juan Gustavo Corvalán, Prometea: Artificial Intelligence to Assist Justice (2021) ⟨https://core.ac.uk/download/pdf/322501055.pdf⟩; Renzo Lavin and others, Use of Artificial Intelligence in Judicial Proceedings (October 2024) ⟨https://fund.ar/en/publicacion/use-of-artificial-intelligence-in-judicial-proceedings/⟩.

[81] Oxford Institute of Technology and Justice, Brazil — Global leader in AI adoption in criminal justice (Accessed: 2025-11-26, 2025).

[82] Victor Habib Lantyer, "The Era of Artificial Intelligence in Law: Brazil in a Global Context" [2023] SSRN Electronic Journal ⟨https://ssrn.com/abstract=4650117⟩.

[83] Jayr Pereira and others, "INACIA: Integrating Large Language Models in Brazilian Audit Courts: Opportunities and Challenges" [2024] arXiv preprint arXiv:2401.05273 ⟨https://arxiv.org/pdf/2401.05273v3⟩.

tation of this tool involves human oversight, as it only suggests an "initial instruction" on the basis of which the decision-maker adjudicates the matter.[84] Researchers also indicate that there is limited risk surrounding the transparency, explainability and accountability of INACIA's decision-making processes.[85]

- VICTOR, a machine learning tool that classifies filed cases, has been adopted by the the Supreme Federal Court.[86]
- Socrates is used by ministers to research precedent, identify similar matters, and analyse appeals for applicable legal provisions and controversies.[87] A tool called Athos is similarly used to identify precedent and divergent positions.[88]
- The Tribunal de Justiça do Rio de Janeiro uses the ASSIS system to draft judicial decisions using GPT-4 based generative models.[89]

3. In China:
    - The Shenzhen Intermediate Court has incorporated LLMs into its workflow to summarise case materials, highlight issues in dispute, analyse relevant precedent, and assist with reasoning and drafting of judgments.[90] In this context, judges draft an initial decision and "the system supports the generation of written justifications and judgment documents."[91]
    - Xiao Zhi, a "robot judge", adjudicates civil cases (e.g., involving consumer credit or private loan agreements).[92]
    - "Smart Courts" are used to look for precedent, recommend applicable laws and regulations, and rectify human errors in the verdict.[93]
    - Courts are using AI systems for criminal sentencing. The 'Intelligent Auxiliary System of Criminal Case Handling', adopted by several courts, analyses

[84] Pereira and others (n 83).
[85] Ibid.
[86] Lantyer (n 82).
[87] Ibid.
[88] Ibid.
[89] Brazil — Global leader in AI adoption in criminal justice (n 81).
[90] John Zhuang Liu and Xueyao Li, "How do Judges Use Large Language Models? Evidence from Shenzhen" (2024) 16(1) Journal of Legal Analysis 235 ⟨https://doi.org/10.1093/jla/laae009⟩; China Daily, "Judiciary embraces AI for efficiency" [2025] China Daily (Accessed: 2025-11-26) ⟨https://www.chinadaily.com.cn/a/202501/02/WS6775eb0aa310f1265a1d8810.html⟩.
[91] Liu and Li (n 90).
[92] Alena Zhabina, How China's AI is automating the legal system (Published by Deutsche Welle (DW),) ⟨https://www.dw.com/en/how-chinas-ai-is-automating-the-legal-system/a-64465988⟩.
[93] Eurasian Times, China's AI-Enabled 'Smart Courts' To Recommend Laws & Draft Legal Docs; Judges To Consult AI Before Verdict ⟨https://www.eurasiantimes.com/chinas-ai-enabled-smart-court-to-recommend-laws-judges/⟩; IndiaAI, "Chinese 'Smart Courts' to recommend laws and draft legal documents" [2025] IndiaAI News (Published online on IndiaAI portal) ⟨https://indiaai.gov.in/news/chinese-smart-courts-to-recommend-laws-and-draft-legal-documents⟩.

facts, identifies issues, and recommends applicable law and sentencing in criminal cases.[94] The High People's Court of Inner Mongolia uses 'Faxin Zhitui' to draft judgments.[95]

4. Estonia has implemented a semi-automated process for issuing payment orders in small claims and maintenance cases, which are treated as judicial decisions for enforcement purposes.[96]
5. In Germany:
    - The Frankfurt District Court uses "Frauke" for assistance in drafting repetitive judgments in class actions suits.[97]
    - The Stuttgart Higher Regional Court uses the Higher Regional Court Assistant (OLGA) to find relevant precedent and produce templates for decisions.[98]

      Both Frauke and OLGA are designed to incorporate human oversight and ensure that decisions are ultimately made by judges, and not the AI tools.[99]
6. In India:
    - The Ministry of Law and Justice issued a press release on the experimental development of SUPACE (Supreme Court Portal Assistance in Court Efficiency) to assist in understanding case facts and searching for precedent.[100]
    - The income-tax department has implemented a Faceless Assessment Scheme with an automated allocation tool to randomly allocate cases to assessing authorities and an automated examination tool that reviews draft assessment orders.[101] Humans are involved at each stage of this process, and are responsible for drafting show-cause notices, income or loss determination proposals, review reports, and

[94]Oxford Institute of Technology and Justice, China — AI deeply embedded in criminal justice system (Accessed: 2025-11-26, 2025).

[95]Ibid.

[96]Ministry of Justice and Digital Affairs, Factsheet: AI Strategy ⟨https://e-estonia.com/wp-content/uploads/factsheet-ai-strategy.pdf⟩; Kai Harmand, "AI Systems' Impact on the Recognition of Foreign Judgements: The Case of Estonia" (2023) 32 Juridica International Law Review 107 ⟨https://doi.org/10.12697/JI.2023.32.09⟩.

[97]Lexology, The Rise of Artificial Intelligence in Legal Practice: Efficiency and Challenges ⟨https://www.lexology.com/library/detail.aspx?g=ffd2a154-171a-4a5c-af5a-97eb9b4d0471⟩.

[98]Hengeler Mueller, The Evolving Role of AI in German Dispute Resolution ⟨https://hengeler-news.com/en/articles/the-evolving-role-of-ai-in-german-dispute-resolution⟩; Experts Institute, Artificial Intelligence: prominent on the agenda of the CEPEJ Plenary Meeting 4 and 5 of December 2023 ⟨https://experts-institute.eu/en/europe-of-justice/cepej-en/cepej-december2023/⟩.

[99]IBM, "Judicial systems are turning to AI to help manage vast quantities of data and expedite case resolution" (2025) ⟨https://www.ibm.com/case-studies/blog/judicial-systems-are-turning-to-ai-to-help-manage-its-vast-quantities-of-data-and-expedite-case-resolution⟩.

[100]Ministry of Law and Justice, "Use of Artificial Intelligence in Supreme Court" (n 17).

[101]Income-tax Act, 1961 1961; Faceless Assessment Scheme 2019; Amendment to the Faceless Assessment Scheme 2021.

orders.[102]

7. In the United States:
    - The Patent and Trademark Office uses AI tools for reviewing patent and trademark applications.[103]
    - The Social Security Administration system uses a tool called Insight to improve the quality of order writing: "At the hearing level, Insight is used to identify weaknesses in draft opinions, ensuring that adjudicators have properly gone through the analysis required by regulations."[104] Insight also provides templates for decisions, and incorporates a set of "quality flags" that examine non-compliance with applicable law and policy and highlight inconsistencies in a draft decision.[105] This indicates a human-in-the-loop, who is accountable for the final decision.

[102]Income-tax Act, 1961 1961.
[103]Engstrom and others (n 43).
[104]Administrative Conference of the United States, *Artificial Intelligence and Regulatory Enforcement* (n 43).
[105]Ibid.

C Annexure C

This Annexure elaborates on the Problem-Solution-Evaluation framework. In Section 2.2, we identified seven problems that might arise when LLMs are used to assist order writing processes. For each identified problem, we provide:

1. A breakdown of technical, design, and systemic solutions
2. Contextualised evaluation metrics

C.1 Non-application of mind

Problem

The probabilistic nature of text generation by LLMs may result in output that demonstrably lacks contextual appropriateness or sound legal reasoning. Should such output be utilised without stringent human critical evaluation, it would represent a significant departure from the application of mind principle, a fundamental requisite within the framework of administrative law.

Solution

Technical Solutions

- *Interface Checkpoints*: Implement mandatory pauses in workflows at designated interface points to facilitate human review.
- *Confidence Score Display*: Exhibit model confidence scores and highlight segments with low confidence for targeted inspection.
- *Dual-Prompt Pipelines*: Use dual-prompting pipelines to generate two distinct LLM outputs, and then identify and flag discrepancies.
- *Functionality Limitation*: Limit system functionality to presenting options rather than finalised recommendations.
- *Constraint Enforcement*: Employ tools, such as *Guardrails AI*,[106] to enforce predefined output constraints and alert reviewers upon violations.

Design solutions

- *Workflow Design for Review*: Design workflows that explicitly require qualified personnel to review, approve, or revise LLM suggestions.

[106] guardrails ai/ guardrails ⟨https://github.com/guardrails-ai/guardrails⟩.

Systemic solutions

- *Role-Based Access*: Restrict editing privileges through role-based access controls to ensure clear accountability.

Evaluation

- *Edit Rate*: The quantum of LLM segments altered or rejected by human reviewers, indicating depth of oversight.
- *Turnaround Time (TAT)*: The duration between an LLM suggestion to human sign-off on it, which can reflect operational efficiency and engagement.
- *Prompt Divergence Rate*: A measure of the frequency with which alternative prompts yield substantially different legal outputs, indicating areas that require human judgment.
- *Coherence Score*: Automated tools, like *RAGAS*,[107] evaluate consistency in references, entities, and terminology, which is useful for pre-screening drafts before human review.

C.2 Black-box problem

Problem

Opaque reasoning paths in LLMs affect the auditability of LLM systems. Specifically, when the logical steps or data points leading to a particular output remain opaque, it becomes challenging to discern the underlying rationale. This might pose challenges to the administrative law requirements for transparency and reasoned orders.

Solution

Technical solutions

- *Chain-of-thought prompting*: Apply methodologies such as chain-of-thought prompting to compel models to explicitly articulate their intermediate reasoning steps.
- *Input Token Influence Identification*: Utilise interpretability techniques[108] like *LIME* (Local Interpretable Model-agnostic Explanations)[109] and SHAP (SHapley Additive

[107] explodinggradients/ragas ⟨https://github.com/explodinggradients/ragas⟩.

[108] The term 'interpretability' refers to the direct understanding of the internal processes of an AI model. The term 'explainability', on the other hand, refers to providing post-hoc explanations for an AI model's output, in line with human reasoning, while its internal mechanisms remain opaque to humans. See, Marcinkevičs and Vogt (n 56)

[109] marcotcr / lime (n 57).

exPlanations)[110] to identify which input token influenced each segment of the generated output.[111]

- *Symbolic Reasoning Systems*: Combine LLMs with symbolic reasoning engines that enforce rule-based logic.[112]
- *Traceability tools*: Incorporate specialised tools that log and record the usage of source documents during the text generation process, to create an auditable trail of information provenance.

Design solutions

- *Visualisation*: Employ visualisation tools to represent attention layers and the relevance of source materials. This can reveal which parts of the input data the model focused on, thereby highlighting the evidential basis for its conclusions.

Systemic solutions

- *Simplified model explanations*: Publish simplified explanations of model functionality to demystify their basic operations for a broader audience, including regulators, adjudicators, and other stakeholders.

Evaluation

- *Clarity rating*: A regulator-assessed rating of the clarity and utility of the model's intermediate reasoning, particularly focusing on the interpretability and usefulness of chain-of-thought outputs.
- *Audit Trail Incidence*: Quantifies the number of inconsistencies or anomalies flagged during periodic audit reviews of the model's operations and its generated outputs. A lower incidence rate indicates higher transparency and auditability.
- *Document Traceability Rate*: Represents the proportion of generated content for which attributable source references can be accurately identified. A high traceability rate is indicative of robust factual provenance, improving transparency.

[110] shap (n 58).

[111] Novel methods like TokenShap extend the concepts of Shapley values to LLMs by *"interpreting LLMs by attributing importance to individual tokens or substrings within input prompts"*.(See Miriam Horovicz and Roni Goldshmidt, "TokenSHAP: Interpreting Large Language Models with Monte Carlo Shapley Value Estimation" [2024] arXiv ⟨https://arxiv.org/pdf/2407.10114⟩) While the scalability of these models are still a concern, there is significant promise that TokenSHAP and similar methods can help interpreting the input to output relationships in LLMs.

[112] Symbolic reasoning systems, and notably neuro-symbolic approaches, have provided valuable contributions to the field of AI. However, it must be acknowledged that the practical application of symbolic reasoning currently remains more speculative than other potential solutions discussed in this paper.

C.3 Potential for bias

Problem

LLMs may perpetuate harmful biases or toxic language from their training data, directly challenging the administrative law principles of the rule against bias and non-arbitrariness.

Solution

Technical solutions

- *Data Preprocessing*: Preprocess training datasets to ensure balanced representation across various demographics and to eliminate or mitigate toxic segments.
- *Bias penalisation*: Employ techniques such as Direct Preference Optimisation (DPO)[113] or Iterative Null-space Projection (INLP)[114] to penalise and thus reduce the generation of biased or undesirable completions during the model's training or fine-tuning phases.
- *Domain-specific content filters*: Deploy specialised content filters designed to flag or entirely block the generation of problematic or inappropriate language within the legal or sectoral domain concerned.
- *Automated Bias Flagging Tools*: Tools such as *Perspective API*[115] can automatically identify and flag content that is potentially biased or toxic.

Systemic solutions

- *Establishment of Legal Fairness Criteria*: Define and establish a comprehensive set of legal fairness criteria that is sector-specific.
- *Mandatory Periodic Benchmarking*: Institute mandatory periodic benchmarking exercises using recognised bias evaluation benchmarks. This ensures continuous monitoring and assessment of the model's tendency to generate biased or toxic output.

Evaluation

- *Bias Flag Rate*: Quantifies the number of segments flagged by automated filters per 1,000 outputs, indicating the system's alertness and sensitivity to potential bias.
- *Override Percentage*: Indicates the share of flagged outputs that human reviewers

[113] Rafael Rafailov and others, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model" [2023] arXiv ⟨https://arxiv.org/abs/2305.18290⟩.

[114] Shauli Ravfogel and others, "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection" (July 2020) ⟨https://aclanthology.org/2020.acl-main.647.pdf⟩.

[115] Perspective API ⟨https://www.perspectiveapi.com/⟩.

subsequently cleared or edited, highlighting the exercise of critical human judgement.

- *Fairness Benchmark Scores*: Measures improvements demonstrated on established bias datasets, such as the Bias Benchmark for QA dataset (BBQ)[116] or the BOLD Dataset and Metrics for Measuring Biases in Open-Ended Language Generation.[117]

C.4 Confabulations

Problem

LLMs have displayed the tendency to confabulate when used in adjudicatory settings. They may invent facts, cite non-existent precedent, or attribute propositions to cases that do not appear in those cases. The tendency of LLMs to confabulate may lead to the violation of the administrative law principles of non-arbitrariness, if their output is not properly verified before it is used.

Solution

Technical solutions

- *Retrieval Augmented Generation*: Use retrieval-augmented generation (RAG) to ground outputs in authoritative legal sources.[118]
- *Post-Generation Verification*: Implement processes for post-generation re-verification of named entities and citations within the output to confirm their validity and accuracy.
- *Legal Knowledge Graph Integration*: Use legal knowledge graphs to validate cited relationships and ensure internal consistency.[119]

Design solutions

- *Mandatory reviewer verification*: Require all output, such as precedent and citations, to be verified by a human reviewer at every stage of the workflow.
- *Watermarking for traceability*: Implement a system to watermark any draft document created or reviewed using LLM-based tools. This ensures clear traceability of LLM

[116] nyu-mll/BBQ ⟨https://github.com/nyu-mll/BBQ⟩.
[117] amazon-science/bold ⟨https://github.com/amazon-science/bold⟩.
[118] Patrick Lewis and others, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" [2020] arXiv ⟨https://arxiv.org/abs/2005.11401⟩.
[119] Hannah and others (n 46).

involvement in content generation or modification, facilitating greater scrutiny.[120]

Systemic solutions

- *Communicate technical limitations*: Through structured training programmes and clear, comprehensive manuals, communicate the inherent technical limitations of the LLM-tool to adjudicators and other potential users. Emphasise the need for detailed and constant scrutiny of LLM-generated output for accuracy and relevance.

Evaluation

- *Secondary LLM "Judges" for Fact-Checking*: Integrate secondary LLM "judges" designed specifically to fact-check the outputs of primary LLMs, providing an additional layer of automated validation.
- *End-to-End Evaluation Tools*: Deploy tools such as RAGAS,[121] which provide comprehensive, end-to-end evaluations for aspects including hallucinations, faithfulness to source material, and relevance of generated content.
- *Hallucination Rate*: Quantifies the number of factually incorrect or fabricated assertions per 1,000 outputs, disaggregated by the type of error (e.g., fabricated information versus irrelevant content).
- *Retrieval Precision@k/ MRR*: Measures the quality and relevance of retrieved legal material used in the output generation process.
- *Self-Consistency Rate*: Determines the proportion of generated variations of a given output that contradict each other. This metric flags internal inconsistencies in generated output.
- *NLI Coherence Check*: Natural Language Inference (NLI) techniques used to detect internal logical contradictions in generated text.[122] Models like DeBERTa-NLI[123] are particularly effective for this purpose, providing an automated means to assess internal consistency.

[120]It is necessary to mention here that watermarking AI-generated content is typically considered a challenging exercise, particularly in terms of its scalability. However, tools like Google DeepMind's SynthID Text propose a scalable framework with little or no impact on quality. (See, Synth ID ⟨https://deepmind.google/models/synthid/⟩). That said, this is more speculative as a solution than the other techniques proposed here.

[121]explodinggradients/ragas (n 107).

[122]Zahra Rahimi and others, "HalluSafe at SemEval-2024 Task 6: An NLI-based Approach to Make LLMs Safer by Better Detecting Hallucinations and Overgeneration Mistakes" (Association for Computational Linguistics June 2024) ⟨https://aclanthology.org/2024.semeval-1.22.pdf⟩.

[123]microsoft/deberta-v3-base ⟨https://huggingface.co/microsoft/deberta-v3-base⟩.

C.5 Metacognition and the Dunning-Kruger Effect in LLMs

Problem

One key element of human understanding of problems and their responses is the ability to understand ones limitations, and their ability thus to learn (meta-cognition). The Socratic Paradox[124] is applicable to human beings as well, often leading to the confidence-competence gap (commonly known as the Dunning Kruger effect in the psychology literature). However, it is expected that an adjudicator will know when their knowledge is limited in a matter, and will presumably attempt to learn more about the subject before writing the order. However, LLMs lack metacognition and there is a confidence-competence gap.[125] This might lead to the LLM being more confident than it should be in output which is limited by its knowledge and the training corpus. It cannot seek or access new points of law or fact when generating its output. This might affect the administrative law principles of application of mind and non-arbitrariness.

Technical solutions

- *Prompt engineering*: Create a system prompt asking the LLM not to answer questions outside of its corpus and mentioning clearly that if it does not find the answer to a question, or finds limited answers to a question, during its retrieval phase, it should generate an answer which clarifies its ignorance. At the same time, the LLM should be prompted to provide references for its answers from within the documents in the corpus.
- *LLM as a judge*: The use of one LLM to retrieve and generate, and then the use of another LLM to judge the answer. The second LLM reads the answer provided by the first and tries to identify issues with the answer, if any.[126]
- *Iterative improvement from feedback*: Luang et al. (2022) provide an experimental set

[124]The Socratic Paradox is that one cannot know what one does not know. It follows then, that if one cannot know what one does not know, how would one understand that they do not know?

[125]Singh and others (n 38).

[126]It must be noted that using LLM-as-a-Judge has some limitations for a variety of reasons. For instance, both LLMs might be trained on a similar corpus, or LLMs might be unable to refute or criticise incorrect propositions due to alignment measures, especially in exchanges involving deep knowledge. (See, Annalisa Szymanski and others, "Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks"March 24–27, 2025, Cagliari, Italy¿1800 (March 24–27, 2025, Cagliari, Italy) ⟨https://dl.acm.org/doi/pdf/10.1145/3708359.3712091⟩) However, this method shows significant promise and performs close to human preferences in certain benchmarks, when a powerful model is used as a judge. (See, Lianmin Zheng and others, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena" [2023] 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks ⟨https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf⟩).

up which showcases that LLMs improve their performance and become more knowledgeable about a subject if they are provided relevant feedback.[127] With our proposed system where officers audit and provide feedback in every step of the process, we expect the confidence-competence gap of the LLM to reduce over time.

Evaluation

- *Closeness Metric*: An automated metric that categorises answers in four quadrants (High Confidence Correct, High Confidence Incorrect, Low Confidence Correct, Low Confidence Incorrect), and then tabulates these for a set of pre-determined tasks.[128]
- *Human evaluation of overconfidence in output*: While the closeness metric is an automated metric based on questions with Correct/Incorrect answers, in order writing in many cases the answers are more qualitative in nature. Since there is no 'total' number of relevant output in this case, human experts will provide a count similar to the Closeness Metric (such as, quadrant per measure of length/page etc.).

C.6 Training corpus

Problem

Legal texts tend to be heterogeneous, noisy, and siloed. Further, legal instruments, such as laws, amendments, cases, and orders, are disparately spread across jurisdictions and time periods, and are continuously evolving. As a result, one of the challenges of developing an efficient legal language model is the insufficiency of readily available training datasets.[129] An LLM trained on a poor corpus is more likely to generate inaccurate and irrelevant output.

Solution

Technical solutions

- *Adaptive Scraping Frameworks*: Use adaptive scraping frameworks, such as *NeuScraper*[130] to acquire data from authoritative legal sources and to perform text cleaning opera-

[127] Jiaxin Huang and others, "Large Language Models Can Self-Improve" [2022] arXiv ⟨https://arxiv.org/abs/2210.11610⟩.
[128] Singh and others (n 38).
[129] See, Hsien-Lin and Cheng (n 40)
[130] Zhipeng Xu and others, "Cleaner Pretraining Corpus Curation with Neural Web Scraping" [2024] arXiv ⟨https://arxiv.org/abs/2402.14652⟩.

tions.[131]

- *Sector-specific pre-training*: Continue pre-training LLMs on sector-specific corpora to refine their understanding and generation of legal language.
- *Structure Entity Extraction and Legal Knowledge Graphs*: Extract structured entities from legal texts and construct legal knowledge graphs for better semantic retrieval.

Systemic solutions

- *Source inclusion*: Define and rigorously enforce quality criteria for the inclusion of primary source materials into the system.
- *Perplexity tracking*: Monitor model perplexity over time to detect instances of overfitting or data drift, which can compromise model performance and generalisability.
- *Legal Retrieval Benchmarking:* Benchmark legal retrieval capabilities using established tools, such as BEIR,[132] to quantitatively measure document-level search relevance and efficiency.
- *Curate sector-specific legal databases*: Regulators must maintain a robust and updated database of all applicable laws, regulations and a linked, annotated precedent bank for their sector.

Evaluation

- *Crawl coverage*: The percentage of designated priority sources successfully scraped and ingested into the system.
- *OCR Error Reduction*: Quantified improvement in text quality post-cleaning, reflecting the efficacy of the data quality pipeline.
- *Validation perplexity*: An indicator of the model's generalisation capability on previously unseen legal text.
- *Retrieval lift*: The measured change in retrieval quality after updates or enhancements to the training corpus.[133]

C.7 Data security and privacy

Problem

[131] Adaptive scraping frameworks have gained wide adoption in recent times. Commonly used tools for data scraping like Scrapy or popular LLMs like Perplexity allow for very targeted scraping for corpus development or referencing of documents for answer generation, respectively.

[132] beir-cellar / beir ⟨https://github.com/beir-cellar/beir⟩.

[133] Yansheng Mao and others, "LIFT: Improving Long Context Understanding of Large Language Models through Long Input Fine-Tuning" [2025] arXiv ⟨https://arxiv.org/html/2502.14644v3⟩.

Documents containing legal pleadings, evidential submissions, and orders may contain sensitive personal data and other classified information, such as propriety information protected by intellectual property laws. The input of such material into LLMs, whether for training purposes or as direct operational input, typically necessitates storage on third-party servers. This practice introduces inherent vulnerabilities, as these external servers may not be adequately protected from the risks posed by potential data breaches or other sophisticated cybersecurity threats.

Solution

Technical solutions

- *Stringent access control*: Implement industry-grade encryption protocols and stringent access controls to safeguard sensitive information.
- *Synthetic supervision-based PII detectors*: Integrate synthetic supervision-based PII (Personally Identifiable Information) detectors[134] to achieve high-recall coverage of sensitive entities across diverse legal inputs. This automates the identification of sensitive information.
- *NLP filters for information masking*: Deploy Natural Language Processing (NLP) filters designed to detect and automatically mask sensitive information, preventing its unintended exposure.
- *Isolated Model Containers*: Isolate model containers from external networks to minimise potential attack vectors.

Systemic solutions

- *On-premise infrastructure*: Operate all LLM infrastructure on secure, on-premise servers to maintain strict control over data residency and access.

Evaluation

- *Unauthorised access attempts*: The volume and frequency of blocked intrusion attempts, reflecting the effectiveness of security measures.
- *Mean Time To Remediation (MTTR)*: The average time taken to patch identified vulnerabilities, indicating the responsiveness of security operations.[135]

[134]Shubhi Asthana and others, "Adaptive PII Mitigation Framework for Large Language Models" [2025] arXiv ⟨https://arxiv.org/html/2501.12465v1⟩.

[135]Imranur Rahman and others, "No Vulnerability Data, No Problem: Towards Predicting Mean Time To Remediate In Open Source Software Dependencies" [2025] arXiv ⟨https://arxiv.org/html/2403.17382v2⟩.

- *Penetration Test Pass Rate*: A measure of the system's resilience against simulated adversarial attempts to penetrate the system.
- *PII Detection Accuracy*: The percentage of sensitive terms correctly identified and masked by automated systems, used to measure the precision of privacy-preserving mechanisms.