WHITE PAPER | JUNE 2025

Legal red teaming, one year in: Reports from the field

> *"As companies roll out chatbots and other generative AI-powered tools, this work is essential to ensure that they don't run afoul of the law in areas like bias, compliance, copyright, and privacy."*
>
> —Bloomberg Law

Legal red teaming, one year in: Reports from the field

Introduction

In our June 2024 white paper, *Legal red teaming: A systematic approach to assessing legal risk of generative AI models*, we presented legal red teaming, a methodology aimed at helping organizations that develop and deploy generative artificial intelligence (GenAI) systems to proactively surface and address legal and compliance risks through adversarial testing. Legal red teaming introduces law and regulation as a grounding framework for adversarial testing of GenAI, complementing more traditional technical and "sociotechnical" attack vectors.

Since the publication of our white paper, the landscape of AI risk assessment and the legal scrutiny surrounding AI has continued to evolve. Consequently, many organizations are no longer asking *whether* to engage in adversarial testing of their GenAI systems, but rather *how* and *when*.

DLA Piper's approach to legal red teaming has been widely recognized by a variety of governmental organizations, educational institutions, and the legal community:

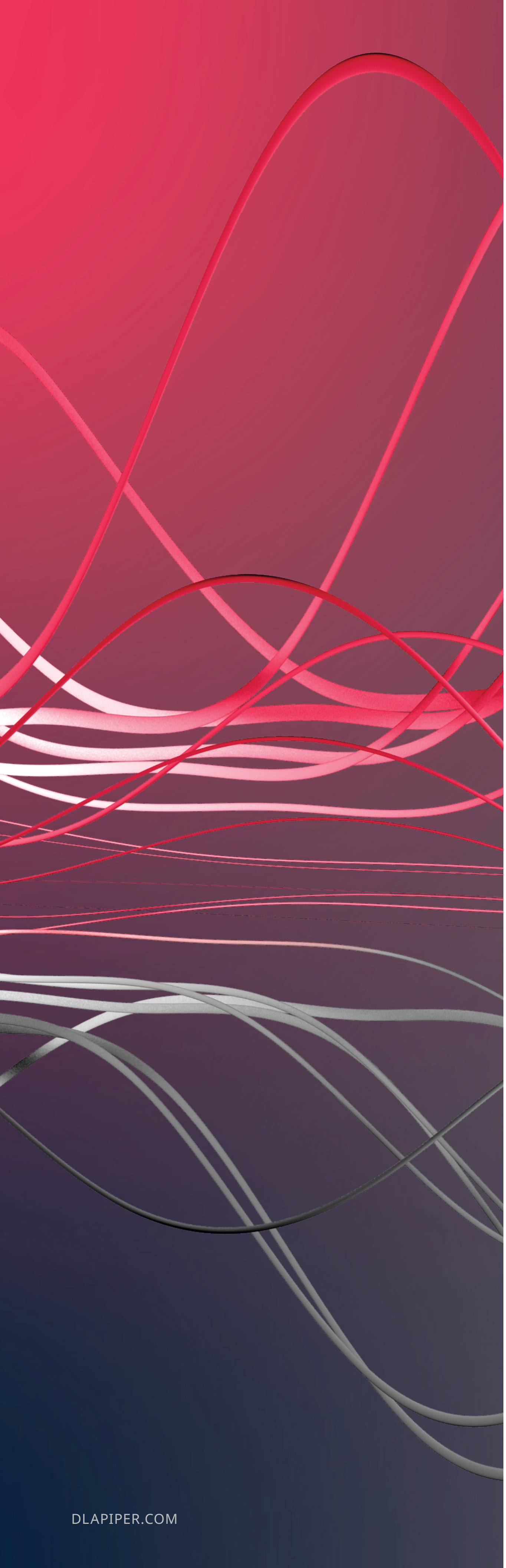
> *"[DLA Piper's] vision for integrating rigorous legal standards with cutting-edge technology provides a roadmap for navigating the complex landscape of AI ethics and regulation."*
>
> —United Nations AI for Good

> *“DLA Piper has taken a forward-thinking view of the intersection of emerging technologies and the law. By embracing direct methods of model evaluation, such as red-teaming, the firm intelligently intersects technological capabilities with rigorous legal oversight.”*
>
> —Dr. Rumman Chowdhury, US Science Envoy, Artificial Intelligence

> *DLA Piper’s legal red teaming reflects “practical efforts ... crucial to deploy such models in a compliant way.”*
>
> —The Oxford Handbook of the Foundation and Regulation of Generative AI (Oxford University Press, forthcoming)

Our legal red teaming approach has been adopted in life sciences, healthcare, retail, and other sectors. The forthcoming *Oxford Handbook of the Foundation and Regulation of Generative AI* noted that our approach is “crucial to deploy[ing] such models in a compliant way.”

Below, we provide an update on legal red teaming, including lessons we have learned from deploying it in practice over the past year. Drawing on real-world engagements and feedback from our clients, we highlight common challenges, share patterns that have emerged across diverse use cases, and propose strategies and best practices for using legal red teaming and technical testing in a complementary way.



Market penetration of GenAI

GenAI models are advancing rapidly; however, they can be unpredictable and present legal, regulatory, and reputational risk. While major hallucinations by GenAI models – such as fabricating legal cases – have garnered attention, a subtler issue is the presence of near truths (referred to as "delusions") and accurate yet unlawful outputs.

This section discusses the extensive adoption of GenAI, some significant issues encountered with these systems, and why trust and safety activities – including red teaming – can help address these challenges.

Adoption of GenAI across sectors

GenAI has become the most rapidly adopted technology in history, outpacing even the personal computer and the internet.[1] This has largely been driven by the technology's broad utility and the ease with which humans can interact with it.

Over the past year, we have seen clients across all sectors expand their use and development of GenAI solutions. In a global survey on the current state of AI, McKinsey found a 17-percent increase in the number of organizations that use AI in at least one business function in a given year (from 55 percent in 2023 to 72 percent in 2024) and a 32-percent uptick in the use of GenAI (from 33 percent in 2023 to 65 percent in 2024).[2]

While the early days of GenAI largely focused on boosting internal productivity within an organization, companies across industries are turning toward development and deployment of tools that can transform the way they do business and interact with the market. For example, in an enterprise survey conducted by Deloitte, companies reported achieving benefits from the use of AI across business activities, including in relation to fraud detection, product development, and innovation.[3]

This trend is not limited to software or technology companies. We have counseled clients on GenAI adoption in the healthcare, retail and consumer goods, finance, and insurance industries, supporting a broad range of use cases for both internal and external applications.

[1] Alexander Bick, Adam Blandin, and David J. Deming, *The Rapid Adoption of Generative AI* 1-2 (Nat'l Bureau of Econ. Rsch., Working Paper No. 32966, 2024), http://www.nber.org/papers/w32966.

[2] "The state of AI: How organizations are rewiring to capture value," McKinsey (Mar. 12, 2025), https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

[3] "Now decides next: Generating a new future," Deloitte (Jan. 2025), https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-gen-ai-q4.pdf.

Current issues with GenAI systems

GenAI systems have significantly advanced in versatility and capability over the past several years, and they are capable of producing high-quality outputs across various tasks. However, meaningful legal and compliance risks associated with GenAI systems persist.

Accuracy remains a concern: Despite significant advances in AI model architecture, training techniques, and data availability, the consistent production of accurate outputs remains a hurdle for GenAI systems and the organizations that develop or deploy them.[4] This challenge becomes increasingly evident when organizations develop systems that shift away from narrowly defined tasks and to more general-purpose applications. For example, even very sophisticated GenAI systems can produce incorrect summaries of technical documentation or complex policies and make incorrect statements regarding product prices or descriptions. Further, these system outputs can blend true and false information in ways that are persuasive yet incorrect, making it difficult for end users to discern whether the output is accurate.[5] GenAI models can hallucinate not only in their outputs but also in their explanations of how they arrived at those outputs.[6] In other words, the apparent explainability of GenAI – its ability to answer questions about how it arrived at a result – can itself be misleading. This phenomenon of *meta-hallucination* makes it even more challenging to identify erroneous outputs.

Accuracy issues are not just a quality or user experience concern – they can also be a business risk. Inaccurate outputs can lead to physical or economic harm to users who rely on GenAI tools, which can give rise to claims under legal liability theories, including those grounded in consumer protection, negligence, and product liability. Defamation cases brought against foundation model makers are just the tip of the spear when it comes to accuracy-related legal claims – and may signal the urgent need for companies to adopt responsible AI practices.

Guardrails are incomplete: Organizations developing and deploying GenAI systems have implemented a variety of guardrails to prevent harmful or undesirable outputs – these include content filters, prompt restrictions, system prompts, and alignment tuning. However, such measures may not be sufficient on their own to reliably prevent outputs that create potential legal, regulatory, or reputational risks.

Guardrails are typically designed to block overtly harmful or offensive content, but they may be inadequate to identify nuanced, subtle, and context-dependent legal and compliance risks. For example, GenAI chatbots may recommend actions that appear reasonable but cause legal or physical harm, such as medication recommendations that are wrong for a particular user, legal guidance that misstates or traverses the law, "do-it-yourself" instructions that lead to safety issues, or financial plans containing hallucinated facts, calculations, or advice leading to unfavorable outcomes. Guardrails attempting to prevent such outputs may be incomplete or erratic – or, where allowed, answers may be inaccurate but so persuasively stated as to appear correct. As GenAI is increasingly coupled with physical devices and agentic architectures that are capable of taking independent action in the real world, the potential for variable or unlawful recommendations to cause harm increases.

Moreover, companies are increasingly modifying open-source or proprietary foundation models, whether through finetuning, prompt engineering, retrieval augmented generation, or other techniques, to specialize or personalize upstream models to their needs. We have found that these companies tend to rely on the original models' guardrails without an assessment of how such modifications could affect their businesses. For example, we have seen modifications to a base model allow users to elicit responses that likely would have been blocked by the original model's guardrails. This is true both with regard to the initial modification of models and modifications made during the development and deployment cycles. It is not always known what modifications will affect existing guardrails. We often hear from data science teams that specific changes have not "touched" the guardrails but find through testing that model performance has changed in that regard.

Technical teams often cannot foresee legal issues: Designing and deploying GenAI systems in ways that minimize legal and enterprise risk while preserving usability is fundamentally a cross-functional challenge. One of the core difficulties is that legal and compliance issues are often unknown, and therefore unforeseeable, to technical teams. This is because legal and compliance risks do not always stem from how a model functions, but rather from contextual factors including how they behave in real-world environments and how users interact with them. As such, a GenAI system may function exactly as designed from a technical standpoint yet still may precipitously increase risk to the organization based on its deployment context. Even in cases in which technical teams are aware of potential risk categories that may be implicated by the GenAI system they are developing, effectively managing the contours of specific legal risks often requires deep legal knowledge.

The cost of adoption and cost of testing may be mismatched: A further challenge that organizations routinely face when developing or deploying GenAI systems is the growing asymmetry between the cost of adopting GenAI and the cost of testing it. GenAI has become remarkably accessible. Pretrained models and open-source frameworks are readily available to organizations for their own use and development. Such accessibility allows organizations to easily experiment and rapidly innovate at relatively low cost. By contrast, the cost of rigorously testing GenAI systems remains high, as GenAI's

[4] See Artificial Intelligence Index Report 2025, at 1170, Stanford Inst. For Human-Centered Artificial Intelligence (Apr. 7, 2025), https://hai.stanford.edu/ai-index/2025-ai-index-report.
[5] Hongshen Xu et al., *Delusions of Large Language Models*, arXiv (Mar. 9, 2025), https://arxiv.org/abs/2503.06709.
[6] *See Id*.

probabilistic, open-ended nature means testing requires time, repetition, knowledge, and appropriate resources. Comprehensive evaluation of GenAI models calls for more than measuring performance against technical benchmarks. It involves stress testing models for varied and unexpected use cases, simulating adversarial prompts, identifying legal and reputational risks, and assessing the systems' robustness across user contexts, both benign and malicious. These tasks require significant knowledge, often involve some level of manual effort, and are iterative – all of which can increase overall cost to the business. The mismatch between the relative ease of adoption and the potential risk profile, and thus cost of testing, can lead to a cognitive dissonance for leadership when it comes to resource allocation. Leadership often tasks their companies with ensuring "safe" or "responsible" AI without adequately funding those mandates. As discussed below, the Department of Justice (DOJ) has called for "proportionate" allocation of resources between investment in, and compliance of, technology. And, while IBM Consulting reports a steady year-over-year increase in the amount of company spending on responsible AI, the relative proportion is still notably low, with responsible AI spend expected to constitute only 5.4 percent of total AI investment in 2025.[7]

The timing between development and testing may be mismatched: Another key pattern we have observed is that legal red teaming, and responsible AI testing more broadly, is often conducted at the last minute, after development decisions have been made and considerable resources expended. This likely arises from AI functions being siloed within many companies, which have not united business enablement and risk management functions in early AI dialogue. An effective enterprise responsible AI program generally combines these value-creation and risk management functions into a working committee that ensures "responsible AI by design." Such approach is often a more cost-effective solution than "AI by design," followed by legal or compliance review shortly before launch. In speaking with corporate professionals, we sometimes find that business functions hope that, by driving innovation forward and delaying the introduction of legal counsel or compliance officers until later, they can make a model or project a *fait accompli* – in other words, create momentum behind a project and "beg forgiveness rather than ask permission." We find such an approach often yields the opposite of the desired outcome. Problems often surface at the last moment – this can lead to rushed legal red teaming and the uncovering of additional issues late in the process, which can delay releases and require costly remediation. A more orderly, organized approach – building responsible AI into the process from the start – can enable business value creation and reduce friction by surfacing issues early, before they are entrenched.

Red teaming may be a key part of the solution

Industry experts and regulators alike have consistently identified red teaming and adequate trust and safety measures as key to managing AI legal and compliance risks. Red teaming is also increasingly recognized as a key risk-mitigating measure for organizational litigation exposure, with some emerging AI-specific laws – such as the Colorado AI Act – explicitly identifying affirmative defense statutory violations that are discovered through red teaming and subsequently remediated.[8]

NIST AI Risk Management Framework

Legal red teaming can play a key role in aligning companies with global risk frameworks, such as the National Institute of Standards and Technology (NIST)'s Artificial Intelligence Risk Management Framework (AI RMF) and Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (GenAI Profile).

NIST is a federal agency within the US Department of Commerce that plays a prominent role in advancing measurement science, standards, and technology to enhance innovation and industrial competitiveness. NIST develops a range of technical standards, guidelines, and frameworks that are widely used across industries to help ensure quality, security, and operational efficiency. These resources can help organizations navigate complex technical and regulatory challenges while fostering best practices in technology development and implementation.

In early 2023, NIST introduced the AI RMF in response to the growing need for effective risk management practices in the realm of AI.[9] In July 2024, NIST published the GenAI Profile as a follow-up to address the specific risk management challenges presented by GenAI.[10] Both frameworks were designed to help organizations identify, assess, and manage potential risks associated with AI systems – ranging from ethical implications to cybersecurity threats. By providing a structured approach to AI risk management, the AI RMF and GenAI Profile support companies in aligning their processes with global standards, thereby helping ensure that AI technologies are deployed responsibly and with due diligence. Both frameworks suggest that organizations implement red teaming to "identify potential adverse behavior or outcomes of a [GenAI] model or system, how they could occur, and stress test safeguards."[11] The GenAI Profile stresses that, "for best results, AI red teams should

[7] "The enterprise guide to AI governance," IBM Institute for Business Value (Oct. 17, 2024), https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-governance.
[8] Colo. Rev. Stat. § 6-1-1706(3).
[9] NIST AI RMF Playbook, NIST (Jan. 2023), https://airc.nist.gov/airmf-resources/airmf/.
[10] Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST (July 26, 2024), https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence.
[11] Nat'l Inst. of Standards & Tech., NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile 50 (2024), https://doi.org/10.6028/NIST.AI.600-1.

demonstrate domain expertise, and awareness of socio-cultural aspects within the deployment context."[12]

By investing in adversarial testing methods, organizations can identify vulnerabilities across their legal and operational domains before such vulnerabilities are exploited. Legal red teaming simulates real-world scenarios, effectively testing whether current policies, controls, and practices are robust enough to withstand evolving risks and align with recognized legal and regulatory frameworks, thereby helping reduce exposure to potential legal and operational pitfalls.

European Union Artificial Intelligence Act

Legal red teaming can also help companies comply with the European Union Artificial Intelligence Act (EU AI Act). Passed by the European Parliament on March 13, 2024 and entered into force on August 1, 2024, the EU AI Act was designed to serve as a comprehensive regulatory framework for AI, classifying systems according to risk levels and operator roles (*eg*, deployer, provider) and establishing corresponding requirements for transparency, accountability, and safety.

For General-Purpose AI Models with Systemic Risk, the EU AI Act mandates that providers "perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting *adversarial testing* of the model with a view to identifying and mitigating systemic risks."[13]

Beyond that requirement, legal red teaming can also simulate potential legal and regulatory challenges that an AI system might face, effectively testing an organization's ability to meet the requirements set forth by the Act. By investing in specialized teams, technologies, and procedures, organizations can continuously evaluate whether their systems comply with evolving legal standards and are prepared to address any enforcement actions or legal uncertainties.

Moreover, legal red teaming's proactive approach can not only help ensure compliance, but also reinforce a company's commitment to ethical and responsible AI deployment. As the EU AI Act aims to foster trust and safety in AI technologies, consistent investment in appropriate red teaming can enable organizations to identify potential risks early, optimize compliance strategies, and integrate best practices across legal and operational domains. This alignment with regulatory expectations could be key in reducing the risk of costly penalties and reputational damage, while also promoting innovation within a framework of ethical responsibility and legal certainty.

DOJ Corporate Compliance Programs guidance

In September 2024, the DOJ updated its Evaluation of Corporate Compliance Programs guidance to address the challenges posed by emerging technologies, and particularly by AI. The revised guidance underscores the DOJ's expectation that companies proactively assess and manage the "emerging risks" associated with AI and other disruptive technologies.[14] It directs prosecutors to evaluate the technology that a company uses to conduct its business, whether the company has assessed the risks associated with using that technology, and whether it has appropriately taken steps to mitigate the risks associated with that technology.[15]

As part of their assessment of such risks, prosecutors will likely also evaluate whether the company has compliance controls and tools in place to assess the accuracy and reliability of the data used by the business operations, and whether the company monitors and tests the technology's functionality and alignment with the company's code of conduct.[16] The guidance emphasizes the necessity for continuous monitoring and testing of AI systems to detect and mitigate potential misuse or unintended consequences, underscoring that companies should invest "the same resources and technology into gathering and leveraging data for compliance purposes that they are using in their business."[17]

Investing proportionally in AI system testing may be crucial for several reasons. Adequate funding enables the development of robust compliance programs that can effectively identify and address AI-related risks, such as data privacy concerns, cybersecurity vulnerabilities, and biases in automated decision-making. The DOJ's guidance highlights the importance of allocating resources to ensure compliance functions have access to relevant data and analytics tools comparable to those used in other business operations. This parity can allow compliance teams to monitor AI systems effectively, ensuring they operate within legal boundaries and company policies.

By dedicating sufficient resources to AI testing and compliance, organizations can not only adhere to regulatory expectations but also foster trust and reliability in their AI-driven initiatives.

[12] *Id.*
[13] Article 55.
[14] US Department of Justice, "Evaluation of Corporate Compliance Programs" at 2 (Sept. 2024), https://www.justice.gov/criminal/criminal-fraud/page/file/937501/dl
[15] *Id.* at 3–4.
[16] *Id.* at 13.
[17] *Id.*

Our work so far

Over the past year, our team has conducted legal red teaming of multiple GenAI tools in development or adoption by Fortune 100 and 500 companies. These organizations are high-profile, global brands launching large language model (LLM)-powered tools with expected user bases of millions. We have performed legal red teaming of a virtual shopping assistant, patient- and physician-facing chatbots, and an individual organization's implementation of an enterprise-wide GenAI assistant, among others.

As set forth in our original white paper, our novel approach to adversarial testing uses the law as a ground truth or gold standard. Our cross-functional legal team engages directly with the GenAI system to elicit outputs that could potentially run afoul of a variety of legal obligations relevant to a given use case, including consumer protection and unfair competition laws, sector-specific regulations (*eg*, Food and Drug Administration, or FDA, regulations; unauthorized practice of medicine), data privacy obligations (*eg*, the Children's Online Privacy Protection Act, or COPPA), intellectual property protections, and a host of other legal and enterprise risk areas.

We also identify and probe areas of potential reputational risks, including harmful or offensive GenAI outputs and guardrail exhaustion. We have elicited controversial political, social, and brand-related statements from our clients' chatbots that, without appropriate guardrails, could have been headline worthy. As noted in our original paper, GenAI is probabilistic and variable – it can answer the same question many ways, and it can respond to a vast swath of open-ended prompts. Thus, red teaming is always about risk reduction, not risk elimination. However, we have seen red teaming tighten model behavior and reduce unwanted outputs when applied.

Technical red teaming remains equally important, and testing for vulnerabilities to malicious technical attacks is a foundational step. While legal red teaming employs some aspects of prompt engineering to elicit responses, it focuses on the legal and reputational content of outputs as opposed to the technical breaking of the system. When paired with technical adversarial testing, this approach can provide a holistic picture of a system's strengths and weaknesses and empower an organization to make risk-informed decisions. These findings can be translated into actionable mitigation strategies and remediation recommendations that are grounded in the law.

Findings from the field

Over the past year, our red teaming exercises have surfaced patterns that are difficult to detect through a single legal red teaming exercise but, through continued engagement with our clients, have surfaced over time. This section outlines issues that we have discovered through our various client engagements and continued work on legal red teaming.

GenAI may produce legally violative outputs: Some GenAI systems frequently produce outputs with the potential to violate laws or regulations. These outputs are not limited to edge cases or misuse scenarios; they often emerge in standard interactions across various contexts. For example, GenAI systems may produce consumer protection issues if they generate inaccurate or misleading content, product liability claims if users rely on outputs to make decisions that cause harm, and regulatory compliance failures if the system produces outputs that conflict with sector-specific rules, from healthcare to insurance to financial services to human resources. Because these systems operate probabilistically and at scale, even low-probability failures can have high-impact consequences – particularly when they affect customers, patients, employees, or other higher-risk categories.

Small changes can have big consequences: Minor changes to GenAI systems can have unforeseen and disproportionate downstream consequences. For example, slight adjustments to how a model frames information can shift an output from a general statement to one that may be interpreted as the provision of medical advice or a product claim. The effect of these changes could go unnoticed during technical testing but carry significant implications under various laws and regulations, such as those related to consumer protection or product liability. Moreover, if the effect of these changes is not adequately addressed, organizational legal risk can increase dramatically for the same tool.

Changes that shouldn't affect guardrails often do: GenAI system guardrails are often more fragile than anticipated. Guardrails are commonly expected to operate independently of architectural modifications such as orchestration logic or exchanging an underlying model in a multi-model system. In practice, we have seen that even purportedly minor architectural changes can lead to guardrail degradation or circumvention. For example, replacing a component model used for summarization, retrieval, or phrasing may inadvertently weaken content filtering or introduce new failure modes that were previously mitigated. This finding could underscore the importance of regular, ongoing system-level testing and evaluation of risk controls – not only at the time of initial deployment.

AI systems built using pre-trained foundation models can act out in new ways: GenAI systems leveraging pretrained foundation models often behave in unpredictable ways once integrated into broader systems. Even when the underlying foundation model has been finetuned or configured to avoid certain types of outputs, those constraints may not hold when additional components are added – for example, prompt chaining and orchestration layers. These system-level modifications can subtly alter how the model interprets inputs or generates outputs, leading to outcomes that raise legal, compliance, or reputational concerns. In practice, models that appear to be risk-managed in isolation can produce unexpected results when deployed as part of a larger, more complex system.

Overly burdensome GenAI system controls could create usability costs: In response to discovering that a GenAI system under development presents heightened legal or compliance risk, organizations often implement additional controls aimed at reducing the risk presented. While companies are encouraged to understand and address legal and compliance considerations, related controls must be proportional to the risk presented and remain context-aware. Implementing overly aggressive, rigid, or expansive controls in an effort to minimize legal and compliance risks could compromise the usability of AI systems, leading to under-communicative models that are unduly restrained to stock answers. We find that red teaming works best as an iterative practice with development teams to strike the right risk/reward balance for a given use case.

Waiting too long to test can be inefficient and costly: Many organizations wait until late in the development process – or even after deployment – to begin testing their GenAI systems for legal and compliance risks. By that point, identifying and addressing issues is often costlier, less effective, and more disruptive to the product under development. In contrast, early-stage testing can allow organizations to surface risks when design decisions are still flexible, mitigation options are broader, and risk managing measures can be implemented efficiently.

Next steps for companies

Based on our experience conducting legal red teaming of internally and externally facing GenAI tools, many pain points that organizations face are process related – and can be avoided by planning for red teaming and other trust and safety activities ahead of time. In this section, we provide an overview of best practices to avoid these pain points and maximize the benefits (and potential cost savings) of legal red teaming.

Build red teaming into the design process: Organizations are encouraged to build legal red teaming into the design and development process early, and plan for phased testing and retesting throughout the project. Organizations that choose to test until immediately before launch can miss key opportunities to mitigate risks up front, often with less expense and less disruption to the overall development timeline.

Unite risk and development functions early: Bringing risk functions into the design and development conversations early may be critical. As discussed above, legal issues are often not foreseeable to teams developing Gen AI systems. When project teams wait to engage legal or compliance until right before product launch, they are often forced to make decisions based on time or budget pressure rather than on benefit to the system's performance and its users as a whole. Upfront buy-in from key stakeholders on appropriate testing, guardrails, and potential legal and enterprise risk areas while a project is still in development can foster collaboration, improve overall performance, more accurately capture budgetary needs for testing and mitigation activities, and potentially avoid difficult last-minute decisions about what can be done versus what should be done from a trust and safety perspective.

Allocate budget for both: Along similar lines, organizations are encouraged to budget for both development and testing (including red teaming) activities up front to the extent possible. Put another way, companies are encouraged to consider appropriate testing and trust and safety measures an indispensable *part of* the design, development, and deployment process. Drawing budget for research and development activities from value centers within an organization while allocating trust and safety activities from cost centers like legal and compliance could lead to a problematic imbalance in design and function. Testing activities may be considered a necessary line item for GenAI development, not an after-the-fact expense viewed – mistakenly – as an unanticipated cost rather than a value-unlocking opportunity.

Set a testing cadence: Legal red teaming and other testing activities should occur at a regular cadence and at key moments in the development lifecycle. GenAI technology is inherently unpredictable and rapidly evolving, and we have seen countless examples of seemingly minor shifts that have caused marked differences in model performance. For this reason, organizations are encouraged to avoid basing their testing on whether a technical change "should" or "shouldn't" change model behavior and instead plan for ongoing testing at key development milestones.

Where we are headed

As we conduct legal red teaming engagements across industries and use cases, we continuously use the insights we gather to advance our approach and meet our clients where they are. Through this ongoing work, we have uncovered opportunities to enhance our legal red teaming services, both in terms of scale and depth. We are developing agentic red teaming methods that simulate diverse user behaviors and legal risk scenarios at scale, and we are forming strategic partnerships that expand the scope of our red teaming services. These advancements can allow us to offer our clients more comprehensive, adaptive, and forward-looking testing strategies. This section outlines some of the ways we are continuing to evolve our red teaming approach.

Multi-agent legal red teaming

According to our findings, human-led legal red teaming remains the most efficient method for surfacing nuanced, context-dependent legal and compliance risks presented by GenAI systems. Attorney red teams bring domain experience and judgment that automated approaches cannot yet replicate. Because of this nuanced and flexible domain understanding, attorneys can craft realistic prompts that elicit responses from GenAI systems, surfacing legal and compliance risks more effectively than in our automated efforts. Further, attorneys are better able to adapt their testing strategies dynamically based on observed behavior and provide more realistic and humanistic interactions. In our experience so far, this makes attorney red teaming well suited for uncovering and identifying system failures and evaluating how those failures will be perceived in real-world legal or regulatory contexts.

However, attorney red teaming has its limitations, most notably with regard to the scale and variation of red teaming strategies. In some cases, legal and compliance risks only emerge under a specific combination of inputs, user behavior, and system configurations. No human red team can exhaustively probe the vast space of possible interactions with a given GenAI system, and some amount of automation will add affordable scale and scope to these efforts. These scaling efforts may help reduce legal and compliance risk blind spots, especially if a GenAI system is deployed to a large user base or in varied use contexts.

To leverage and scale attorney capabilities, DLA Piper is developing a multi-agent legal red teaming framework. This approach allows our attorney red team members to extend the scope of their red teaming efforts by generating parameterized red teaming prompts through the assistance of LLM-based agents at the attorney's direction. Using this multi-agent approach, we are able to simulate diverse legal and compliance probes at scale while leveraging, not replacing, nuanced attorney domain knowledge and judgment. By combining the depth of attorney knowledge and experience with the breadth of automated exploration, we are building a more comprehensive and scalable legal red teaming capability for organizations deploying GenAI systems.

Collaboration with traditional red teaming methods

To advance a holistic approach, we believe in the integration of traditional technical red teaming methods – focused on security, robustness, and misuse – with legal red teaming, which probes systems for compliance, legal, and reputational vulnerabilities. With this in mind, DLA Piper has an alliance with Scale AI to combine these complementary forms of red teaming in a single engagement. Our collaboration with Scale AI provides a fulsome analysis which simulates both intentional misuse and benign, real-world uses that may result in legal, compliance, or security risks, providing organizations a robust evaluation of risk management, and thus business enablement, when harnessing the value of GenAI systems.

About us

DLA Piper is a global law firm with lawyers located in more than 40 countries throughout the Americas, Europe, the Middle East, Africa and Asia Pacific, positioning us to help companies with their legal needs around the world.

For more information

To learn more about DLA Piper, visit dlapiper.com or contact:

Danny Tobey
Partner and Global Co-Chair and Chair of DLA Piper Americas AI and Data Analytics practice
+1 214 743 4538
danny.tobey@us.dlapiper.com

Ashley Allen Carr
Partner
+1 512 457 7251
ashley.carr@us.dlapiper.com

Barclay Blair
Senior Managing Director
AI Innovation Lead
+1 212 335 4709
barclay.blair@us.dlapiper.com

Karley Buckley
Associate
+1 713 425 8421
karley.buckley@us.dlapiper.com

Sean Fulton
Of Counsel
+1 214 743 4572
sean.fulton@us.dlapiper.com

Kyle Kloeppel
Associate
+1 916 930 3253
kyle.kloeppel@us.dlapiper.com

100+

Lawyers, data scientists, coders, and policymakers focused on AI worldwide

Artificial Intelligence Global Market Leaders Spotlight
Chambers Global 2025

AI Leading Lawyer
Danny Tobey
Band 1
Chambers USA 2025

Innovation in Generative AI Strategy
Financial Times 2024

Innovation in New Services to Manage Risk
Financial Times 2024

Best Law Firm Use of AI
American Lawyer 2024

Innovative Lawyers in Technology
Financial Times 2023

Top AI Lawyers
Danny Tobey
Business Insider 2022

This publication is for general information only. The information presented is not legal advice, and your use of it does not create a lawyer-client relationship. All legal matters are unique and any prior results described in this publication do not guarantee a similar outcome in future matters. DLA Piper is a global law firm operating through DLA Piper LLP (US) and affiliated entities. For further information, please refer to dlapiper.com. Attorney Advertising. Copyright © 2025 DLA Piper LLP (US). All rights reserved.