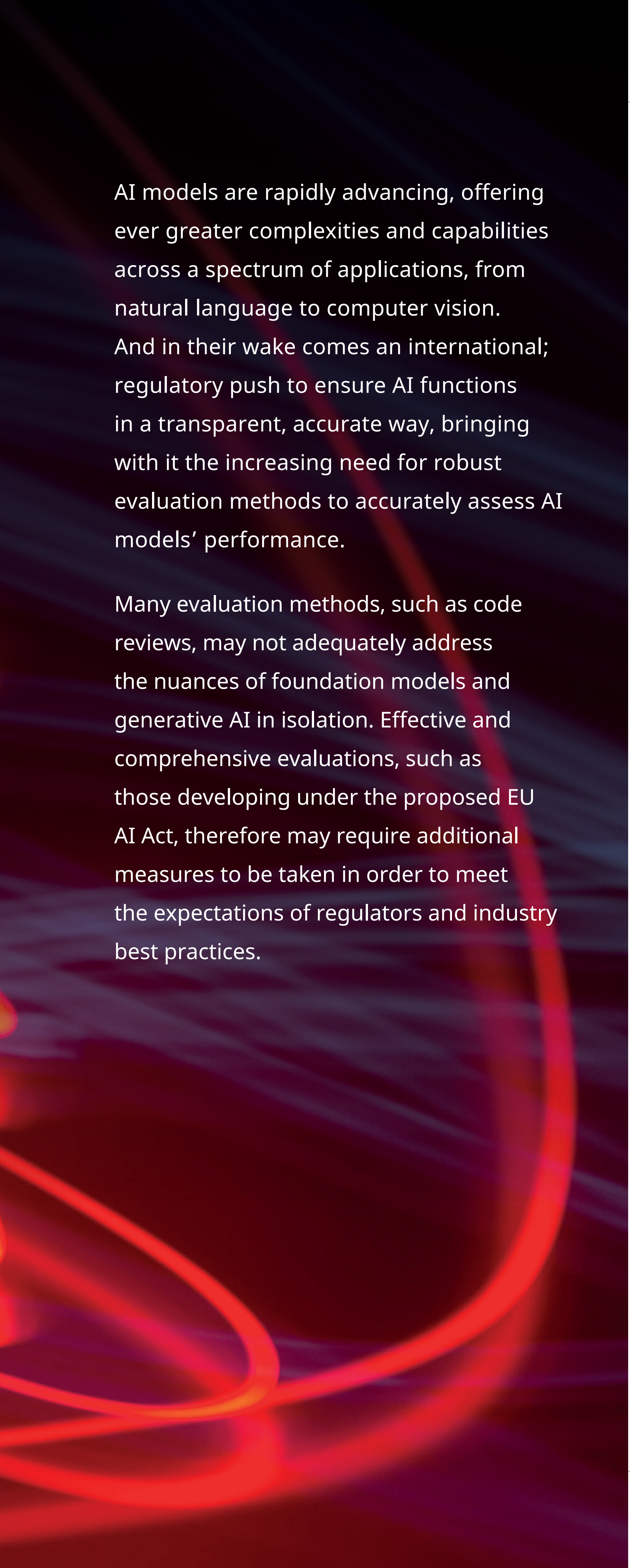
Red teaming: A practical guide to testing AI models

DLA PIPER



AI models are rapidly advancing, offering ever greater complexities and capabilities across a spectrum of applications, from natural language to computer vision. And in their wake comes an international; regulatory push to ensure AI functions in a transparent, accurate way, bringing with it the increasing need for robust evaluation methods to accurately assess AI models' performance.

Many evaluation methods, such as code reviews, may not adequately address the nuances of foundation models and generative AI in isolation. Effective and comprehensive evaluations, such as those developing under the proposed EU AI Act, therefore may require additional measures to be taken in order to meet the expectations of regulators and industry best practices.

What is red teaming?

One alternative (or complimentary) way to evaluate the accuracy and technical robustness of an AI model is red teaming. This method involves deliberately probing an AI model to test the limits of its capabilities. This may be done manually, by teams of individuals (similar to video game testing in which bug searches are carried out) or by pitting AI models against each other.

There are two primary goals to this process:

- uncover specific vulnerabilities and identify priority areas for risk mitigation and
- utilize red team attempts as input datasets to foster the development of safer systems.

Particular focus during earlier stages of developing an AI model should be given to identifying risks, vulnerabilities, emergent unintended biases, and undesirable behaviors that may arise in the model's performance. The discovery of any vulnerabilities should not be viewed negatively; this helps to strengthen the AI system by pinpointing its capabilities and areas for improvement. This ensures continuous improvement and a safer user experience. It is also a crucial step in identifying system vulnerabilities that may impact the resilience of the models and the hardware on which they rely.

Can I benefit from red teaming?

Red teaming may be applied to any situation that relies on accuracy and reliability of the tools used. For example:

- **Healthcare**: AI continues to move into how we care for patients and treat illnesses. Red teaming can be used in this context to ensure that systems are able to diagnose or test with accuracy and act in a way that can be relied on. It may be the case that successful red teaming could be used as evidence of successful safety testing when seeking to achieve safety certifications.
- **Foundation models**: Red teaming may allow providers of foundation models to see how their system behaves when faced with undesirable prompts, such as requests for scam emails, creation of discriminatory media, and hateful content. Creators will be able to tailor their guardrails to ensure that users cannot get around current restrictions in order to use their platforms for this form of content.

- **Editorials**: Fake news continues to be a battle faced by content creators and information-sharing sites. The use of red teaming in this context could be used to test AI models designed to restrict the sharing and distribution of false information by probing whether certain phrases or characters are being deployed to sneak past its protocols.

As indicted above, these are merely examples of areas in which red teaming may be beneficial. Red teaming can be used in all contexts and across all industries.

Practical considerations for a successful red team evaluation

The process of red teaming can seem unwieldy and complex without appropriate parameters and initial planning. To prepare for a future evaluation, consider the following high-level factors to better understand the questions that need to be asked before a successful Red Teaming testing phase.

DETERMINE YOUR EVALUATION CRITERIA

To understand the results of the process, it is necessary to clearly define the criteria used to evaluate the outputs provided. These could include:

- **Accuracy and factual correctness**: Is the output accurate or factually correct, given the parameters of the evaluation? For example, is a summary of a complex document accurate?
- **Relevance to input**: Is the output or analysis relevant to the proposed input? For example, when using an AI model for detecting fake news, did it correctly determine whether information was fake news, or did it alert to legitimate information?
- **Contextual coherence**: Is the output coherent, based on the context? Does a generative AI model provide an appropriate answer based on the context of the information that can be inferred by the data provided?
- **Logical flow**: Does the output logically flow throughout the use of the AI model? For example, does a medical tool recommend an appropriate treatment, after discovering the patient has an allergy to certain antibiotics?
- **Adversarial robustness**: Does the AI stand up to deliberate attempts to break or confuse the model? For example, does it understand when paradoxical information (such as basing future calculations off of inaccurate base numerical data) is provided and can it respond with an appropriate output?
- **Resistance to ambiguity**: Is the AI able to hand vague or ambiguous information within inputs and provide coherent and accurate results? For example, is an AI customer service bot able to make sense of customers' problems with limited information and direct them to the correct resources?

DETERMINE YOUR METHODOLOGY

After determining your intended evaluation criteria, it is necessary to understand how you plan to perform the Red Teaming process. Considerations could include:

- **Man or machine**: Will the process be performed by individuals, or by pitting AI models against each other, or by a combination of both?
- **Hammer or scalpel**: If another AI model is being used in the process, will it be a large model with vast quantities of data or will it be a specific model with limited or tailored data that can be used as inputs? For example, adversarial AI systems can look for a specific issue (sexism, toxicity) or can be more generally applicable.
- **Task master**: What specific task is the AI model that is being red teamed intended to accomplish, and how do you intend to test its limits on this basis?
- **Keeping score**: At each instance when an output is produced, how is this to be recorded, and what specific data points are intended to be used for analysis?

DETERMINE HOW YOUR RESULTS WILL BE INTERPRETED

Now that the evaluation criteria and testing methods have been established, it is important to consider how to interpret the results and score the AI model on that basis. This could include:

- **Strengths and weaknesses**: Provide a breakdown of the strengths and weaknesses of the model, highlighting what it was successful in responding to and what forms of inputs highlighted issues within the system.

- **Numerical value**: Assign certain evaluation criteria with numerical values and determine whether the final score of the model meets the intended level of performance, taking note of low scoring factors that could be rectified at a later date.
- **Critical risk analysis**: Identify any critical risks or vulnerabilities that must be addressed, such as zero-day exploits, which are attacks that take advantage of security vulnerabilities that do not have fixes in place or processes that allow the AI model to behave in a way that is in direct contravention of its set guidelines and use policy.

DETERMINE FREQUENCY OF RED TEAMING AND CHANGES TO PARAMETERS

Performing the process on one occasion may be insufficient. As AI models and technology continue to develop, new exploits and risks may emerge. Therefore, it is recommended to conduct red teaming and other forms of analysis on a routine basis. The frequency of such analysis should depend on the use case and potential risks associated with its use and is subject to internal processes and evaluation procedures.

Red team ready

At DLA Piper, our integrated AI & Data Analytics team stands at the intersection of law and technology, comprising top-tier lawyers, data scientists, analysts, and policymakers leading AI development and deployment. We are a pioneering blend of lawyer-data scientists that seamlessly combine legal acumen with technical depth.

For more information on how to evaluate your AI systems, including foundation models and generative AI, and to keep up to date on the emerging legal and regulatory standards, please contact any of the authors, and visit DLA Piper's Focus page on Artificial Intelligence.

For more information

Find out more about our firm at dlapiper.com or contact:

Dr. Sam Tyner-Monroe
Managing Director
T +1 202 799 4522
sam.tyner-monroe@us.dlapiper.com

Bennett B. Borden
Partner, Chief Data Scientist
T +1 202 799 4357
bennett.borden@us.dlapiper.com

Christopher Cullen
Of Counsel
T +1 215 656 2492
christopher.cullen@us.dlapiper.com

Coran Darling
Law Clerk (Admitted in England & Wales)
T +1 212 335 4703
coran.darling@us.dlapiper.com

This publication is for general information only. The information presented is not legal advice, and your use of it does not create an attorney-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 © 2023 DLA Piper LLP (US). All rights reserved.
Oct 23 2023 | MRKT0004336_v5LM