

## Comments to the Consultation Paper on Guidelines for Responsible Usage of AI/ML In Indian Securities Markets, dated 20 June 2025

On 20 June 2025, the Securities and Exchange Board of India released the “[Consultation Paper on guidelines for responsible usage of AI/ML In Indian Securities Markets](#)” (hereafter “**Consultation Paper**”).

The Consultation Paper sets out the landscape of use cases of AI prevalent in the Indian Securities Market. It also discusses the various frameworks of responsible AI promulgated by various Indian ministries and international organisations; key among them are the NITI Aayog and IOSCO. The Working Group constituted to contemplate the guidelines for the use of AI/ML in the Securities Market undertook desk research, stakeholder consultations to propose a suite of measures along the five principles of responsible AI, i.e., Model Governance, Investor Protection-Disclosure, Testing framework, Fairness and bias, Data Privacy and Cyber Security measures.

In this response, we present our comments to the Consultation Paper. We divide our comments into two parts. **Section A** provides specific input into the scope of principles set out in section 5 of the Consultation Paper. Building on our research on responsible and trustworthy AI in the digital lending sector, we present four considerations that may expand the scope of principles set out in Section 5 and **further strengthen investor rights**. Specifically, we suggest (1) mandating the institution of internal AI Ethical Review Boards that can align the design, development, deployment and performance of AI systems with ethical, legal, and regulatory norms, (b) laying greater emphasis on managing external counterparties through proportionate access rights, security checks and other safeguards, (c) contemplating a right of the investor to contest the decision of AI systems, and, (d) emphasise the creation of board-approved fairness policies to minimise bias and discrimination in the working of the AI systems. **In Section B**, we set out **a list of best practices** that operationalise the key requirements of responsible and trustworthy AI. The list of best practices is designed as a self-assessment questionnaire that allows financial service providers to score the strength of their existing safeguards and gauge their distance from the frontier of best practices. These may be considered to further enrich Annexure B of the Consultation Paper.

---

### Section A: Input to Sections 5.1-5.5

**1. Mandating organisations to institute internal AI Ethical Review Boards** can strengthen model governance, increase trustworthiness of the AI and enhance accountability. An internal AI Ethical Review Board or Committee is a group or body formed within an organization that sets the ethical, legal and regulatory contours within which AI development and deployment must occur. It provides guidance, oversight, and expertise on ethical considerations related to AI and ML, while ensuring AI systems and applications are developed, deployed, and used in a responsible, fair manner and aligned with ethical principles and societal values.<sup>1,2</sup> Several organisations have already instituted such boards.<sup>3</sup> Wipro has an internal taskforce dedicated to the design, development and deployment of responsible AI. The taskforce emphasizes embedding “*transparency, accountability, privacy, non-discrimination, accessibility, participation, right to redress, and human oversight for AI initiatives in order to protect human rights.*”<sup>4</sup>

Such internal boards bear the promise of folding in the interest of all stakeholders. When constituted correctly, i.e. with participation from ethicists, data protection and customer protection experts, independent members, these boards have the potential to prevent investor-protection from becoming an

---

<sup>1</sup> <https://dvararesearch.com/wp-content/uploads/2025/04/Responsible-and-Trustworthy-AI-Framework.pdf>

<sup>2</sup> <https://www.bigdataframework.org/knowledge/the-role-of-ai-ethics-boards-navigating-the-ethical-landscape-of-artificial-intelligence/>

<sup>3</sup> <https://economictimes.indiatimes.com/tech/artificial-intelligence/ai-ethics-boards-coming-up-fast-in-indian-tech-majors/articleshow/112814960.cms?from=mdr>

<sup>4</sup> <https://www.wipro.com/content/dam/nexus/en/sustainability/pdf/Human-Rights-Policy.pdf>

afterthought and indeed fold in the interest of stakeholders at the model design/development/selection stage.<sup>5</sup>

We recommend that the scope of the principle of Model Governance be expanded to include a mandate of instituting internal AI Ethical Review Boards.

**2. Management of external counterparties is an essential aspect of model governance.** The Consultation Paper recognises the participation of several third parties in the design, development and deployment of AI systems. These third parties may also need to access the organisation's environment. This calls for differentiated access controls and proportionate responsibilities, commensurate with the role of the third party vendor. This consideration also implications for allocation of liability. While the regulator holds the regulated entity liable for any deficiencies, it is likely that the parties distribute liabilities contractually among themselves. Highlighting the roles, responsibilities, defining access rights of third parties may help allocate liabilities efficiently and allow for proper indemnification. Some considerations here include instituting security audits of third parties, instituting API Management Systems to prevent abuse and ensure fair usage; instituting Monitoring and Logging Systems that notify administrators of any anomalies or deviations from expected behaviour, and instituting access management systems to regulate the access offered to third parties.

This aspect of managing external counterparties may be considered in the scope of the principle of model governance.

**3. Investor's right to contest algorithmic decision can bolster investor protection and incentivize explainability of AI systems.** Several leading manifestos of responsible AI such as that of the OECD,<sup>6</sup> the UNESCO and the European General Data Protection Regulation<sup>7</sup> accord the right to contest AI decisions. This right can be imagined as a natural extension of the first principle of natural justice, i.e., *the right of the other party to be heard (Audi Alteram Partem)*.<sup>8</sup> While the principle of Investor Protection-Disclosure already emphasises a grievance redress mechanism, the explicit provisioning of a right to contestation will provide the desired heft to the grievance redress mechanism. The recognition of this right will also have implicate the need for transparency and explainability in and of the logic of the AI systems thus, further strengthening the salient recommendations already made in the Consultation Paper.

We propose the principle of Investor Protection-Disclosure to include investor's right to contest automated decisions.

**4. Instituting a board-approved fairness policy and using it to monitor the performance of the model through its lifecycle.** We welcome the caution against bias and discrimination that may filter through AI systems. The debate on bias and discrimination in AI is now well settled. It is appreciated that humans can be biased and discriminatory but what raises concerns about bias and discrimination in AI is the scale of its operation and the difficulty in holding the AI system accountable. The commitment against bias and discrimination can be further strengthened by requiring organisations to institute board-approved fairness policy to define bias and identify its presence, shortlist metrics to measure bias and approve policies to tackle it. The Department of Telecom for instance, has set out a standard for fairness certification of AI. It provides a three-step template to (i) identify bias, (ii) determine the threshold for metrics, and (iii) a methodology for testing for bias.<sup>9</sup> A board-approved fairness policy necessitates defining bias, which is a difficult term to define and measure. Establishing the definition and metrics, proves to be a helpful ex-ante tool for bias avoidance while also doubling up as an ex-post tool for bias correction. Some important components of the fairness policy include guidelines for: (i) defining bias, (ii) debiasing training data, (iii) identifying metrics for measuring bias, (iv) establishing systems for tracking the metrics, and (v) model repairation i.e., tackling bias after it has been detected.

---

<sup>5</sup> <https://link.springer.com/content/pdf/10.1007/s43681-023-00409-y.pdf>

<sup>6</sup> <https://www.oecd.org/en/topics/ai-principles.html>

<sup>7</sup> <https://gdpr-info.eu>

<sup>8</sup> <https://gblrsccpl.in/2023/09/18/supreme-court-reads-audi-alteram-partem-into-rbis-master-directions-on-frauds-a-new-dilemma-for-the-banks/>

<sup>9</sup> [https://www.tec.gov.in/pdf/SDs/TEC%20Standard%20for%20fairness%20assessment%20and%20rating%20of%20AI%20systems%20Final%20v5%202023\\_07\\_04.pdf](https://www.tec.gov.in/pdf/SDs/TEC%20Standard%20for%20fairness%20assessment%20and%20rating%20of%20AI%20systems%20Final%20v5%202023_07_04.pdf)

### Section B: Suite of best practices

In this section we set out a list of best practices that appear necessary to achieve responsible and trustworthy AI. These best practices are curated to mitigate the risks introduced or amplified by the introduction of AI in financial services. In the table below, the column titled ‘RTAI Processes’ sets out the essential requirements of responsible and trustworthy AI systems. The column titled ‘FSP Response’ sets out the different actions that the financial service provider may take to address these requirements. These responses are set out from the least desired to the best practice, awarding a score of 0 to the least desired practice and a score of 8 to the best one. These graded responses allow for a richer, non-binary understanding of the safeguards that FSPs may already have in place to ensure their AI systems behave responsibly. Not all risks are equally profound and therefore, weightages across the responses may vary. This checklist was initially designed by Dvara Research, PwC India and FACE (Fintech Association for Customer Empowerment) to address the needs of digital lenders in India, but we imagine it to be relevant across financial services. This questionnaire is intended to serve as a distance map, in that, it allows financial service providers to measure the distance of their existing practices (reflected in their scores) from the frontier of best practices (the highest scoring practice for each process).

| <b>RTAI (Responsible and Trustworthy AI) Processes</b>   | <b>FSP Response (Score 0-8 )</b>   |
|--|--|
| To what degree are there capabilities in place to audit models?  | 0: Model Audits are done On-demand<br>1: Regular frequency (at least annual audits) with checks on model accuracy (done internally)<br>2: Regular frequency (at least annual audits) with checks on model accuracy (done by a 3rd party)<br>3: Regular frequency (at least annual audits) with checks on model accuracy, logs, access (done internally)<br>4: Regular frequency (at least annual audits) with checks on model accuracy, logs, access (done by a 3rd party) |
| Do you have clearly defined roles in your organization, pertaining to the different aspects of model life-cycle?   | 0: No involvement<br>0: Maker<br>0: Checker<br>6: Approver<br>8: All three (Checker, Maker and Approver)   |
| To what extent have you put in measures to ensure that your employees are aware of the pillars of RAI? Have you considered conducting trainings/awareness programs for the same? | 0: Never conducted<br>1: Once as part of an onboarding exercise<br>2: Annually - only for Makers<br>3: Annually - for Makers and Checkers<br>4: Annually - for all of the above and the users of the AI system including Makers, Checkers and Approvers  |

|  |  |
|--|--|
| <p>To what degree are there HITL (Human in the Loop) governance protocols for the model outcomes? Indicative of presence of human underwriters/model stewards.</p>   | <p>0: No oversight in the models decisions<br/> 0: Can only monitor, but cannot flag out.<br/> 0: Can flag out model decisions, but cannot override them<br/> 6: Can override only in limited scenarios<br/> 8: Can override as per their judgement</p>  |
| <p>How do you keep track of model performance?</p>   | <p>0: Not keeping track<br/> 0: Check for accuracy only during development<br/> 0: Check for accuracy during production and benchmark with development accuracy<br/> 6: #2, along with an established model governance cycle to share feedback with development team on the results from production<br/> 8: Fully automated near-real time MLOps in place to account and correct for deviations from the production outputs.</p>   |
| <p>What kind of controls have you put in place to be able to address any disruptions/unwanted situations (example: disruption could be due to underlying algorithm being unresponsive/server/infra related disruptions)?</p> | <p>0: No mitigation plan in place<br/> 1: Take down the AI model without any ability to redirect the load<br/> 2: Take down the AI model and redirect load to a human agent<br/> 3: Take down the AI model and redirect load to a BRE<br/> 4: Take down the AI model and redirect to a parallel instance of another AI model</p>   |
| <p>To what degree are there security processes in place to prevent unintended use of AI?</p>   | <p>0: No monitoring, controls, or safeguards exist.<br/> 0: Basic controls and monitoring with limited ability to detect or prevent unintended use. E.g. access control (manual and request based)<br/> 0: Standard controls and monitoring are implemented with the capability to detect and prevent the unintended usage. E.g. access control (including underlying codebase - manual and request based), logs monitoring<br/> 6: Advanced controls and monitoring implemented with the capability to detect and prevent unintended usage. E.g. access control (completely automated), real-time logs monitoring and intent identification<br/> 8: Highest degree of controls and monitoring, proactive threat detection and blocking access before an incidence occurs.</p> |
| <p>To what degree are there controls in place to disable the algorithm if there is an issue or unintended use of the AI solution is detected?</p>  | <p>0: No controls in place to disable the algorithm<br/> 1: Disabling the algorithm typically requires significant manual intervention.<br/> 2: Some automated alerts and partial automation in the disabling process.<br/> 3: High level of automation in detection and disabling, with manual override capabilities.<br/> 4: Immediate, automated shutdown procedures are in place, with manual override capabilities for additional safety.</p>   |

|   |  |
|---|--|
| <p>To what degree are there fairness and governance considerations in place in the training data? Mention any steps taken to ensure that the data being used is representative of the entire population demographics</p>  | <p>0: No considerations in place<br/> 0: Removal of data pertaining to demographic data features such as region/religion/ethnicity/gender etc.<br/> 0: Masking of demographic data<br/> 6: Down-sampling/Gen AI based data augmentation of demographic data<br/> 8: Identification and treatment of features that could result in some sort of implicit bias (job, city etc.)</p>  |
| <p>To what degree is there a governance policy to define the fairness rules in the models? Have you addressed any potential for implicit bias in your model?</p>  | <p>0: No policies in place<br/> 1: Policies only catering to identification of known sources of discrimination<br/> 2: Policies only catering to identification of underlying sources of discrimination<br/> 3: Policies only catering to identification and addressal of known sources of discrimination<br/> 4: Policies only catering to identification and addressal of underlying sources of discrimination</p>   |
| <p>To what extent have you employed tests/mechanisms to check for bias in your model's outputs? Have you employed any fairness metrics as part of your model evaluation process? How do you toe the line between improving accuracy and addressing bias in cases of conflict?</p> | <p>0: No metrics/mechanisms to measure bias<br/> 0: Basic mechanisms (example: use test control ratios) to measure bias on strata of sensitive population cohorts (covering 2-3 cohorts)<br/> 0: Basic mechanisms (example: use test control ratios) to measure bias on strata of sensitive population cohorts (covering all relevant cohorts)<br/> 6: Advanced mechanisms (example: indexing and grading) to measure bias on strata of sensitive population cohorts (covering 2-3 cohorts)<br/> 8: Advance mechanisms (example: indexing and grading) to measure bias on strata of sensitive population cohorts (covering all relevant cohorts)</p> |
| <p>To what extent are there processes to cure the bias if it is indeed found in the models?</p>   | <p>0: Unable to detect bias<br/> 0: Able to detect but not address bias<br/> 0: Able to detect, but can only make improvements in the next roll out of the model<br/> 6: Able to detect and address bias only in periodic intervals<br/> 8: Able to immediately detect and address bias in the model in production</p>   |
| <p>To what degree are there capabilities in place to provide visibility to regulators on explainability? Explain if any measures used to describe how your model is able to arrive at an outcome in your statement? Do these meet conditions set by regulators?</p>               | <p>0: No capabilities to provide visibility to regulators on model explainability.<br/> 1: Basic measures, such as high-level explanations, are used to describe model outcomes.<br/> 2: Standard measures, such as feature importance and decision trees, are used to describe model outcomes.<br/> 3: Advanced measures, such as SHAP values, LIME, and interpretable surrogate models, are used to describe model outcomes.<br/> 4: Best-in-class measures, including model documentation, interpretable machine learning techniques, and transparent reporting, are used to describe how the model arrives at an outcome.</p>                    |

|   |   |
|---|---|
| <p>To what degree are there data privacy considerations to protect PII? Do you have different techniques based on the level of sensitivity/confidentiality that has to be maintained? Is this in line with the practices suggested by various data governance frameworks?</p> | <p>0: The lender has no data privacy considerations to protect PII<br/> 0: Basic techniques are used, but there is little differentiation based on sensitivity/confidentiality levels<br/> 0: Techniques such as encryption and access controls are used based on the level of sensitivity/confidentiality.<br/> 6: Advanced techniques, including data masking and anonymization, are used based on sensitivity/confidentiality levels.<br/> 8: Best-in-class techniques, including differential privacy , are used based on detailed levels of sensitivity/confidentiality, completely aligned with standards suggested by leading data governance frameworks</p> |
| <p>To what degree are there data encryption provisions in place to secure the data? Have you put in considerations for data security in transit?</p>  | <p>0: No encryption practices being followed<br/> 1: Generalized masking of data in development environment<br/> 2: Encryption based on levels of sensitivity of all data in the development environment<br/> 3: Generalized masking of data in all stages from development environment and in transit<br/> 4: Employed advanced techniques such as differential privacy to protect individual data points with end to end data encryption( in development, transit and processing)</p>   |
| <p>To what extent is there a special level of security access to interact with the production models?</p>   | <p>0: No measures available<br/> 0: Access given to group IDs(3rd party)<br/> 0: Access given to group IDs(internal)<br/> 6: Access given to named IDs (3rd party) and reviewed periodically<br/> 8: Access given to named IDs (internal) and reviewed periodically</p>   |
| <p>To what degree is there a process in place to delete data for those who wish to be forgotten?</p>  | <p>0: No mechanism for individuals to request data deletion.<br/> 0: An ad-hoc process for handling data deletion requests, often handled on a case-by-case basis.<br/> 0: A somewhat structured process for handling data deletion requests, with defined steps.<br/> 6: Individuals can easily request data deletion through a clear and accessible procedure, such as a dedicated online form or customer service contact.<br/> 8: A comprehensive, highly efficient, and user-friendly process for handling data deletion requests.</p>   |
| <p>To what extent have you taken into consideration the various guidelines set by the likes of RBI and DPDP to protect your customer's rights?</p>  | <p>0:No policies or practices are in place to protect customer rights as per these guidelines.<br/> 1:Partial compliance with regulatory requirements, with significant gaps<br/> 2:Implementation of practices to protect customer rights is somewhat structured and consistent.<br/> 3:Implementation of practices to protect customer rights is thorough and consistent across the organization.<br/> 4:Implementation of practices to protect customer rights is exemplary, with continuous monitoring, improvement, and proactive engagement with regulatory bodies.</p>   |
| <p>To what degree have you employed feedback systems across the</p>   | <p>0:No mechanisms for developers or users to provide feedback.<br/> 0:Basic mechanisms may exist for developers or users to provide feedback, but they are not comprehensive or systematically used.</p>   |

|  |  |
|--|--|
| <p>MDLC? This includes both feedback by developers in the process as well as those of the users</p>  | <p>0:Users can provide feedback, but the system may not be fully integrated or consistently used.(needs some sort of manual intervention)<br/>6:Well-integrated user feedback systems that regularly collect, analyse, and incorporate user feedback.<br/>8:Feedback from both developers and users is systematically collected, analysed, and incorporated into the development process in a continuous manner with help of surveys, focus groups, and real-time feedback tools..</p>   |
| <p>To what measure have you put in facilities that give the user the right to interact with their data? Is the process transparent enough for the user to be able to get a basic understanding of the model outcomes?</p>  | <p>0:Users have no way to understand or question the decisions made by the models.<br/>0:Limited transparency, with users having little insight into how their data is used or how decisions are made.<br/>0:Moderate transparency, but language used in the explanations may still be technical and not fully clear to all users.<br/>6:Users can easily access, update, and delete their data, and receive clear, understandable explanations of model outcomes.<br/>8:Users have full control over their data, including access, updates, deletion, and the ability to see how their data is being used with intuitive, user-friendly interfaces and proactive communication</p>  |
| <p>To what degree have you incorporated various tools and techniques to ensure that your data collection is done in a manner that is clear and unambiguous to the user? Please state if any consent forms are used or user awareness programs being conducted to support your rating</p> | <p>0:No tools or techniques to ensure clarity in data collection.<br/>1:Basic tools and techniques such as a basic consent form or awareness program that is not comprehensive, detailed or user friendly<br/>2:Consent forms are used and provide essential information but may not cover all aspects in detail.<br/>3Detailed and user-friendly consent forms and interactive awareness programs are used, clearly explaining data collection, usage, and user rights.<br/>4:Detailed, clear, and easily understandable consent forms and awareness programs are used/conducted, with explicit descriptions and examples of data collection, usage, and user rights. This is done along with feedback mechanisms that are put in place to gather user input and continuously improve data collection transparency and practices.</p> |
| <p>To what degree is there use of explainable modelling techniques?</p>  | <p>0:No model explainability techniques are being used<br/>0:Basic explanations by using feature importance scores<br/>0:Give insight into the model's results by using a proxy/surrogate model<br/>6:Use model explainability techniques like SHAP/LIME/PDP<br/>8:In addition to point 2 and 3 emphasis made on continuous improvement of explainability, including user-friendly visualizations and clear, non-technical explanations.</p>   |
| <p>To what extent do you put in formal documentation of any and all processes in the MDLC</p>  | <p>0:There is no formal documentation of processes in the MDLC.<br/>0:Processes are ad-hoc and not standardized.<br/>0:Documentation is somewhat detailed but may lack consistency and comprehensiveness.<br/>6:Documentation is detailed, consistent, and regularly updated.<br/>8:Documentation is exhaustive, consistently maintained, and regularly reviewed and updated.</p>  |
| <p>What steps have you taken to prevent your Gen AI application from</p>   | <p>0: No safety measures in place<br/>1: Basic filters for harmful content<br/>2: Regular updates to prevent harmful outputs<br/>3: Comprehensive safety protocols and monitoring</p>  |

|  |  |
|--|--|
| generating harmful content?  | 4: Real-time intervention and correction mechanisms<br><b>NA</b>   |
| How do you mitigate the risk of model hallucination or false information generation in financial advice when interacting with customers?                                       | 0: No mitigation measures<br>0: Basic checks for hallucinations<br>0: Regular monitoring for accuracy<br>6: Advanced verification protocols<br>8: Real-time validation, optimized prompt engineering and model parameter checks<br><b>NA</b>   |
| Do you have any mechanisms in place to segregate external untrusted content from user prompts  | 0: Complete reliance on user discretion without any automated checks or alerts.<br>1: Sporadic identification and manual handling of untrusted content.<br>2: Reliance on manual checks or user input to identify untrusted content.<br>3: Regular reviews and updates of content segregation protocols.<br>4: Technical guardrails are implemented to segregate and/or denote when external or untrusted content is used in a prompt to the Generative AI solution.<br><b>NA</b>  |
| Do you have any protocols that ensure Gen AI applications are thoroughly tested before putting them up in production environment?  | 0: All testing and configuration changes are directly conducted in the production environment.<br>1: Limited or ad hoc use of a non-production environment for testing.<br>2: A basic testing environment exists, but it is not fully isolated from production.<br>3: A sandbox environment is established and used for most testing and validation of the AI solution. Isolation from the production environment is implemented but not fully enforced in all cases.<br>4: A sandbox environment, isolated from the production environment, is used to test the AI solution as part of a phased rollout across the enterprise, and also is used to validate any configuration changes as part of the change management.<br><b>NA</b>  |
| Have you put in any measures to involve human oversight over Gen AI model interactions to protect your application against malicious attacks (prompt attacks, jailbreaks etc.) | 0: No red team assessments or equivalent security evaluations are performed on AI solutions.<br>1: Ad hoc or informal security reviews are conducted instead of structured red team assessments.<br>2: Limited red team assessments are performed, focusing on high-risk components of AI solutions.<br>3: Assessment results are used to inform security improvements, though feedback loops may not be fully optimized.<br>4: Red team assessments are performed on the AI solution(s) prior to deployment within the Organization's environment. These assessments simulate real-world attacks to identify vulnerabilities and potential exploitation methods, providing critical insights into the security posture of the AI solutions and helping to mitigate risks before they can be exploited by malicious actors.<br><b>NA</b> |
| How well prepared is your organization with regards to tackling any  | 0: No IR exercises or training are conducted with a focus on AI solutions.<br>1: Basic IR training is conducted, but it is not tailored to Gen AI-specific incidents.  |

|   |   |
|---|---|
| <p>attacks/incidents/threats that may find its way to your systems through AI/Gen AI applications? Have you taken any steps to inform your team of these threats?</p> | <p>2:IR exercises include general cyber incident scenarios, with some focus on Gen AI solutions.</p> <p>3:AI-specific IR exercises and training are conducted periodically, covering a range of potential incidents.</p> <p>4: Incident Response (IR) cyber exercises and training is conducted, specifically tailored to AI solution(s). These exercises and training sessions will simulate AI-specific cyber incidents, such as model poisoning or adversarial attacks, to test and improve the incident response plan (IRP). Regularly conducting these AI-focused activities ensures that the response team is well-prepared to handle real-world incidents involving AI, thereby enhancing the overall security and resilience of the AI solutions.</p>   |
| <p>What is your guide to governing use of 3rd Party solutions/applications?</p>   | <p>0:No security due diligence is conducted on third-party AI solutions.</p> <p>0:Procurement is largely ad hoc, with limited oversight or assurance mechanisms.</p> <p>0:Procurement involves input from various stakeholders, but lacks a centralized governance team for assurance.</p> <p>6:A dedicated team oversees procurement, but assurance processes may occasionally lack thoroughness or consistency.</p> <p>8:AI third-party or external solutions (e.g., applications, APIs, plug-ins) undergo security due diligence in alignment with Information Security requirements. Contingency processes are in place to handle third-party incidents. The procurement of these solutions is managed and governed by a team, where the main task is to provide assurance of the technology being used.</p>  |
| <p>What mechanisms have you put in place to ensure that your Gen AI application is able to identify and tackle adversarial prompts?</p>                               | <p>0:No adversarial robustness techniques are implemented during model training.</p> <p>1:Limited use of adversarial robustness techniques, often exploratory or experimental.</p> <p>2:Basic adversarial robustness techniques are applied selectively to key models.</p> <p>3:Adversarial robustness techniques are integrated into the training process for most but not all models.</p> <p>4:Adversarial robustness techniques are utilized during model training to enhance the model's resilience against adversarial attacks (e.g., data randomization, adversarial training, federated learning, distillation, ensemble methods).</p> <p><b>NA</b></p>  |
| <p>Does your organization have established data classification protocols to ensure the proper handling and protection of sensitive information?</p>                   | <p>0:No data classification protocols are in place. The organization does not categorize or handle sensitive data differently from other types of data.</p> <p>1:There are informal or ad-hoc practices, but no formal protocols are documented or followed.</p> <p>2:Basic data classification protocols exist. Some categories of data are defined, and there are guidelines for handling them, but they are not comprehensive or consistently applied across the organization.</p> <p>3: Well-defined data classification protocols are in place. Most types of data are categorized, and there are clear procedures for handling and protecting sensitive information. These protocols are generally followed by employees.</p> <p>4:Comprehensive and robust data classification protocols are fully implemented. All data is systematically classified, and there are stringent</p> |

|   |   |
|---|---|
|   | <p>procedures for handling, accessing, and protecting each category of data. Regular training and audits ensure compliance and continuous improvement.</p>  |
| <p>Please indicate the deployment environment for your Large Language Model (LLM)</p>   | <p>0: The LLM is deployed on a standard public cloud service, accessible via the internet with basic security measures.<br/> 1: The LLM is deployed on a public cloud with enterprise-level security features and settings, offering convenience with additional security configurations<br/> 2: The LLM is hosted on a public cloud platform but accessed through a secure VPN, adding an extra layer of security beyond standard cloud measures.<br/> 3: The LLM is hosted on a private cloud with dedicated resources, providing substantial control and enhanced security measures tailored to your needs.<br/> 4: The LLM is hosted and managed entirely within your organization's infrastructure, offering maximum control and security customization.<br/> <b>NA</b></p>                                  |
| <p>Do you have any specific benchmarks are utilized to assess the performance of the LLM in comparison to others in the market?</p> | <p>0: No specific benchmarks are utilized to assess the model's performance against others in the market.<br/> 1: Occasionally, informal or ad-hoc benchmarks are used, but there is no structured approach to performance evaluation.<br/> 2: Some well-known benchmarks are used intermittently, providing a basic level of comparison with other models.<br/> 3: A comprehensive set of established benchmarks is regularly employed to assess performance, offering a clear understanding of the model's standing in the market.<br/> 4: A robust and systematic benchmarking process is consistently applied, using industry-standard metrics and leaderboards to ensure the model is competitively evaluated and continuously improved.<br/> <b>NA</b></p>  |
| <p>Is there a documented schedule for updating the data used in the model?</p>  | <p>0: There is no documented schedule for updating the data, and data versioning is not managed, leading to potential inaccuracies and inconsistencies.<br/> 1: Data updates are made on an ad-hoc basis with minimal documentation, and data versioning is inconsistently applied, resulting in occasional discrepancies.<br/> 2: There is a basic schedule for data updates, and data versioning is applied sporadically, which helps maintain some level of accuracy and consistency.<br/> 3: Data updates follow a documented schedule, with regular versioning practices in place to ensure data accuracy and consistency across model iterations.<br/> 4: A well-defined and strictly followed schedule for data updates is in place, complemented by a robust data versioning system that ensures high</p> |

|   |  |
|---|--|
|   | accuracy and consistency, facilitating seamless tracking and auditing of data changes.   |
| What measures are in place to ensure that the data remains current and representative over time?                                  | <p>0: No measures are in place to ensure data currency or representativeness, leading to outdated or biased data being used in the model.</p> <p>1: Limited efforts are made to update data or assess its representativeness, often relying on occasional reviews or updates without a structured process.</p> <p>2: There are some established procedures for periodic data updates and checks for representativeness, but these are not comprehensive or consistently applied.</p> <p>3: A regular and structured process is in place for updating data and evaluating its representativeness, utilizing feedback and analysis to guide improvements.</p> <p>4: A proactive and systematic approach is employed to ensure data remains up-to-date and representative, including continuous monitoring, stakeholder engagement, and the use of advanced tools and methods to assess and enhance data quality and diversity.</p>   |
| What strategies are employed to prevent overfitting, and how is it ensured that the model effectively generalizes to unseen data? | <p>0: No strategies are in place to prevent overfitting, and there is no focus on ensuring the model generalizes well to unseen data.</p> <p>1: Basic strategies like early stopping or dropout are occasionally used, but there is limited evaluation of the model's generalization ability.</p> <p>2: Common techniques such as regularization, cross-validation, and dropout are applied regularly, with some testing on validation datasets to assess generalization.</p> <p>3: A range of strategies, including data augmentation, regularization, and ensemble methods, are systematically applied. The model's ability to generalize is evaluated across multiple validation datasets with detailed analysis.</p> <p>4: A robust and holistic approach is employed, integrating advanced techniques like transfer learning, hyperparameter tuning, and extensive cross-validation. The model undergoes rigorous testing on diverse and unseen datasets, with continuous monitoring and adaptation to ensure optimal generalization.</p> |

Table B.1: A checklist of best practices for responsible and trustworthy AI, *Source: [Responsible and Trustworthy AI in Digital Lending: From Principles to Practice, Dvara Research, PwC India](#).*<sup>10</sup>

<sup>10</sup> Practices coloured in amber are relevant to GenAI only.