AI Vendor Security and Safety Assessment Guide

This guide helps you evaluate AI vendors by providing essential questions about security, safety, privacy, and ethics. Use it to prepare for vendor meetings and collect important information before signing any contracts.

Version 1.0 - April 2025

Important Considerations Before meeting with a vendor

Define Business Needs

Clearly understand the business problem the AI system is intended to solve and the expected benefits.

Assess Alternatives

Confirm that AI is the most appropriate solution. Sometimes simpler, existing alternatives can meet needs at lower cost and risk.

Prioritize Assessment

Focus your most thorough assessment on high-risk scenarios and systems rather than applying the same level of scrutiny to every vendor.

Customize

Tailor the questions below based on your organization's specific risk tolerance, industry regulations, data sensitivity, and the intended AI application.

Follow-up

This guide facilitates information gathering about the vendors. A separate, formal threat and risk assessment should always be conducted on the solution before going into production.

Questions for Your Internal Business Client

To be asked internally before meeting with the vendor :

Impact Assessment

Is the intended use case likely to significantly impact individuals' rights, safety, health, finances, or access to essential services? Does it involve high-stakes decisions?

Data Sensitivity

Will the AI system process, store, or generate personally identifiable information (PII), protected health information (PHI), financial data, confidential corporate information, or other sensitive data types?

Criticality

How critical is this AI system to core business operations? What is the impact of potential failure or unavailability?

AI Governance & Oversight

- Does your organization have a dedicated AI Governance body or committee overseeing AI development, deployment, and risk management?
- What ethical guidelines, principles, or codes of conduct govern your AI development and deployment?
- Do your developers, data scientists, and relevant personnel receive formal training on responsible AI principles, security, and privacy?

Regulatory Compliance & Standards

General Security Frameworks

What general security frameworks and standards does your organization adhere to (e.g., SOC 2, ISO 27001, NIST CSF)? Provide evidence of certification/attestation if applicable.

AI-Focused Standards

Does your organization or product specifically follow any AI-focused standards or frameworks (e.g., ISO/IEC 42001, NIST AI Risk Management Framework)?

Global AI Regulations

How do you track and ensure compliance with evolving global AI regulations relevant to your service and our use case (e.g., EU AI Act, state-level AI laws)?

Ethical AI Practices & Bias Mitigation

- How do you identify, test for, and mitigate potential biases (e.g., demographic, societal) in your AI models and the data used to train them?
- What processes are in place to ensure fairness and equity in the AI system's outputs and operation?
- Have you performed specific testing to detect and prevent the generation or propagation of toxic, harmful, or discriminatory content?
- What is the provenance of the primary datasets used for training the model(s)? How do you ensure data sources are appropriate and ethically sourced?

Third-Party Validation

Independent Assessment

Has your AI system or security posture been assessed or audited by an independent third party? Can you share the results or attestation reports (e.g., SOC 2 report, penetration test summary)?

Client-Conducted Testing

Are clients permitted to conduct their own independent security assessments or penetration tests (under agreed terms)?

Data Handling & Processing

- Describe the data protection and privacy measures implemented specifically for data processed by your AI systems (inputs, outputs, feedback).
- How is our organization's data (including prompts, inputs, uploaded files, and generated outputs) logically and/or physically segregated from other clients' data?
- Will our organization's data, prompts, or outputs be used to train or fine-tune your models? If so, under what conditions, and can we opt-out?
- Do you review client prompts or data for monitoring, quality assurance, or model improvement purposes? If yes, what controls govern this access?

Data Residency & Sovereignty

Storage Location

Where is customer data (including inputs, outputs, metadata, backups) processed and stored geographically?

Location Restrictions

Can clients specify or restrict data storage locations to comply with data residency requirements (e.g., within the EU, Canada)?

Geographic Access

Can access to or use of AI features be restricted based on geographic regions?

Data Security Controls

Encryption Standards

What encryption standards are used for data in transit and at rest?

Encryption Keys

Can our organization utilize its own encryption keys (Bring Your Own Key - BYOK) for enhanced data protection?

Technical Controls

What specific technical controls exist to detect and prevent users from uploading PII, sensitive, or proprietary information inappropriately?

Data Retention & Deletion

- Describe your data retention policies for client data, including default durations, deletion procedures upon request, and procedures upon contract termination.
- Are data retention policies configurable by clients (e.g., setting shorter retention for prompt/output data)?
- Can users (or administrators on their behalf) permanently delete their specific inputs and outputs (including uploaded files) from the system?
- Upon contract termination, what is the process and timeframe for ensuring the permanent deletion of all client data from primary systems and backups?

Data Ownership

> Who owns the outputs generated by the AI system based on our inputs? Clarify data ownership and intellectual property rights.

Robustness & Adversarial Resistance

Attack Prevention

What safeguards are implemented to detect and prevent prompt injection attacks, jailbreaking attempts, and malicious adversarial inputs?

Training Pipeline Protection

How do you protect the integrity of your model training pipeline against data poisoning attacks?

Adversarial Training

Has the model undergone adversarial training or red-teaming exercises to improve its resilience against misuse and unexpected inputs?

Input Sensitivity

How resilient is the model to significant output changes caused by minimal alterations in input or training data (sensitivity)?

Content Safety & Filtering

Technical Measures

What technical measures (e.g., filters, guardrails) and testing processes are in place to detect and prevent the generation or processing of harmful, illegal, inappropriate, or biased content in both inputs and outputs?

Client Configuration

Are these safety measures configurable by clients?

Vulnerability Management & Misuse Prevention

Vulnerability Management

- What processes exist to identify, track, and remediate vulnerabilities specific to the AI model(s) themselves (beyond standard software vulnerabilities)?
- How frequently do you test the model for security vulnerabilities, performance degradation, and safety alignment?

Misuse Prevention & Hallucination Mitigation

- What processes and technical controls exist to prevent intentional or unintentional misuse of the AI system?
- What techniques or processes are employed to detect and mitigate AI "hallucinations" (confidently incorrect or fabricated outputs)? How are users informed about the potential for such outputs?

Explainability & Model Documentation

Explainability & Interpretability

Describe the level of explainability or interpretability your AI system provides. How can users or auditors understand how the model arrived at specific outputs, decisions, or recommendations (if applicable to the model type)?

Model Documentation

- How are model limitations, intended use cases, and potential risks documented and communicated clearly to customers?
- Do you provide a model scorecard or factsheet detailing performance metrics, training data characteristics, and known limitations?
- How are significant model updates, changes in functionality, or performance shifts communicated and deployed to customers?

Auditability & Monitoring

Audit Logs

Are detailed audit logs available capturing user interactions (e.g., prompts, sessions), system actions, and administrative changes? How long are these retained?

Performance Monitoring

What performance monitoring is in place for the AI systems (e.g., latency, uptime, query success rates)?

Model Drift

How are model drift (performance degradation over time) and significant shifts in output behavior detected and managed?

Reporting

What reporting is available to customers regarding system usage, performance, and security events?

User Notification

Is there a clear notice provided to end-users indicating that they are interacting with an AI system and that outputs are AI-generated?

Operational Security & Infrastructure

Access Control & Authentication

- How is access to the AI systems and related data controlled within your organization (principle of least privilege)?
- What authentication and authorization mechanisms are available for client access to the AI service (e.g., SSO, MFA, role-based access control)?
- Is it possible to implement granular controls or policies when using API functionalities?

Session Management & Monitoring

- Are there functionalities for real-time monitoring of user sessions?
- What measures prevent unauthorized content exfiltration or abuse during active sessions?

Nth-Party & Supply Chain Risk

Nth-Party Components

What nth-party components, services, or foundation models are used within your AI system?

Risk Management

How do you assess, manage, and monitor the security risks associated with these dependencies (including AI supply chain risks)?

Data Sharing Oversight

What oversight exists for any data sharing required with these nth parties? Do you maintain visibility into your critical supply chain components?

Incident Response & Business Continuity

- Describe your incident response plan, particularly for security breaches involving AI systems or client data. How and when would clients be notified?
- How do you ensure business continuity and disaster recovery in case of service disruption? What are your RTO/RPO targets?

Implementation & Control

Human Oversight & Intervention

- What mechanisms allow for human intervention, review, or override of AI-driven processes or decisions, especially in high-risk applications?
- Does your model development lifecycle include a human-review or oversight stage before models are deployed into production?

Configuration & Disablement

- Can the AI functionality be entirely disabled for specific users, groups, or the entire organization if needed?
- Are security guardrails and usage policies configurable by the client administrator?

Final notes

When possible, incorporate these questions into your existing Third-Party Risk Management (TPRM) framework rather than creating a separate process.

A thorough evaluation of vendor responses is critical, requiring careful assessment against your organization's risk appetite, compliance obligations, and specific use case requirements.

Comprehensive documentation of all responses provides a foundation for informed decision-making and establishes accountability in the vendor relationship.

While vendor assurances offer important insights, they must be reinforced through formal contractual obligations, independent verification where feasible, dedicated threat risk assessments of the solution, and continuous monitoring practices.

These combined measures create a robust framework for managing AI risks effectively, ensuring appropriate governance and control throughout the technology's implementation and operational lifecycle.