AI Governance

Vendor Evaluation Cheat Sheet

Customer-Facing, Enquiry Management & Appointment Booking Domain

---

Prepared by Jake Stennett
March 2026

Version 2.0

Introduction

This document provides a structured set of questions for evaluating AI vendors in the customer-facing, enquiry management, and appointment-booking domains. It is designed for use during procurement, RFP evaluation, or vendor due diligence.

The questions are organised into thematic sections covering technology, data governance, security, cost, compliance and operational resilience. A glossary of key terms is included at the end for quick reference.

Section A: Technology & Architecture

A1. What is the underlying AI technology, and how are updates managed?

Determine whether the vendor uses Amazon Bedrock, Azure AI Foundry, OpenAI, an open-source model (e.g. Llama, Mistral), or a proprietary solution. Understand who controls model versioning and what the update/rollback process looks like.

- What LLM(s) does your solution use and what version(s)?
- How and when are model updates deployed? Do you notify customers in advance?
- Can we pin to a specific model version to avoid unexpected behavioural changes?
- What is your rollback procedure if a new model version causes issues?

A2. Is this a walled garden, or is our data shared across tenants?

Establish whether the model learns from your data, whether your data is shared across customers/tenants, and how isolation is maintained.

- Is our instance single-tenant or multi-tenant?
- Is our data used to train, fine-tune, or improve the model for other customers?
- If using a third-party foundation model (e.g. OpenAI), what data-sharing agreements are in place with that provider?
- Where is our data stored at rest and in transit?

A3. How does the solution handle the probabilistic nature of AI responses?

AI is probabilistic rather than deterministic (unlike traditional IVR flows). Understand how the vendor measures, manages and mitigates variability.

- What is your typical accuracy/correctness rate for customer-facing responses?
- Have you conducted structured testing by feeding identical inputs and evaluating output variance?
- What guardrails are in place to prevent hallucinations, off-topic responses, or inappropriate content?
- How do you handle edge cases or ambiguous customer queries?
- What temperature or sampling settings are used, and can we configure them?

A4. What is the architecture for integrating with our existing systems?

Understand what data the AI pushes to and pulls from your CRM, scheduling tools, and other systems.

- What APIs and integration methods does the tool use (REST, webhooks, native connectors)?
- Does it support real-time or batch data exchange with our CRM?
- What happens if an integration endpoint is unavailable, how does the system degrade gracefully?
- Can we audit all read/write operations the AI performs against our systems?

Section B: Security & Data Governance

B1. What identity verification and access controls are in place?

If the AI integrates with CRM or customer record systems, robust ID&V (Identity & Verification) is essential.

- What IDVA procedures are in place before the AI can access or modify customer records?
- How do you safeguard against prompt injection attacks designed to bypass verification and access additional records?
- Do you enforce deterministic flows (rather than AI-driven flows) for identity verification steps? (Recommended , see note below.)

Note: Best practice is to hand off to deterministic flows for all ID&V steps. The AI should never be the sole gatekeeper for accessing sensitive customer data.

B2. How is personally identifiable information (PII) handled?

- Is PII redacted or masked in conversation logs and training data?
- What data retention policies apply to conversation transcripts?
- Can customers request deletion of their interaction data (right to erasure)?
- Is PII encrypted at rest and in transit?

B3. What audit trails and logging are maintained?

- Are all AI decisions, actions, and system interactions logged?
- Can we access full conversation transcripts for quality assurance and dispute resolution?
- How long are logs retained and in what format?
- Are logs tamper-proof and available for regulatory audit?

B4. What is your approach to prompt injection and adversarial input?

Beyond ID&V, prompt injection is a broader security concern across all AI interactions.

- What input sanitisation or filtering is applied before queries reach the model?
- Have you conducted red-teaming or penetration testing specifically targeting prompt injection?
- Do you have documented incident examples and how they were mitigated?

Section C: Cost & Commercial Model

C1. How is pricing structured?

Understand whether costs are per enquiry, per seat, per token, or a fixed subscription , and who bears the risk of cost overruns.

- Is pricing per enquiry, per token, per user, or a flat subscription?
- If token-based, are costs passed directly to us or absorbed by the vendor?
- What mechanisms are in place for cost control (e.g. budget caps, throttling, alerts)?
- What protections exist against token abuse (e.g. users or bots deliberately generating excessive traffic)?
- Are there different cost tiers for different query complexity (simple FAQ vs. multi-turn booking)?

C2. What are the total cost of ownership considerations?

- Are there additional costs for integration, onboarding, training, or support?
- What is the cost model for updates, new model versions, or feature enhancements?

- What notice period applies for pricing changes?

Section D: Performance & Operational Resilience

D1. What performance monitoring is in place?

Identify how the vendor detects issues in live operation, including model drift if the system learns from interactions.

- What real-time monitoring dashboards or alerting systems are available?
- How do you detect and measure model drift over time?
- What SLAs do you offer for uptime, response latency, and resolution accuracy?
- Do you provide regular performance reports or scorecards?

D2. What happens when the AI gets it wrong or cannot handle a query?

- What is the human escalation pathway when the AI cannot resolve a query?
- How quickly does handoff to a human agent occur?
- Is there a confidence threshold below which the AI automatically escalates?
- Can we configure escalation rules and thresholds ourselves?

D3. What is your disaster recovery and fallback strategy?

- If the AI system goes down entirely, what fallback is provided (e.g. revert to traditional IVR, queue to human agents)?
- What is the RTO (Recovery Time Objective) and RPO (Recovery Point Objective)?
- Have you tested failover scenarios and can you share results?

D4. How do you handle bias, fairness and quality assurance?

- Have you tested the model for bias across protected characteristics (age, ethnicity, gender, disability)?
- What ongoing QA processes are in place for response quality?
- Do you have a process for customers to flag and report problematic AI responses?

Section E: Compliance, Ethics & Transparency

E1. Demonstrate compliance with applicable legislation.

The vendor should evidence compliance with all relevant AI and data protection legislation in the jurisdictions where their customers operate.

- GDPR (General Data Protection Regulation) , data processing, consent, right to erasure
- UK Data Protection Act 2018
- EU AI Act , risk classification, transparency obligations, conformity assessments
- Any sector-specific regulations (e.g. FCA for financial services, CQC for healthcare)

E2. Are customers informed they are interacting with AI?

- Is there a clear disclosure to the end user that they are speaking with an AI system?
- Is this disclosure compliant with the EU AI Act transparency requirements?
- Can the customer opt out and speak to a human at any point?

E3. What is your responsible AI framework?

- Do you have a published responsible AI policy or ethical AI principles?
- Is there an internal AI ethics board or review process?

- How do you handle requests from regulators or public authorities regarding AI decisions?

E4. What are the vendor lock-in risks?

- Can we export our configuration, training data, and conversation history if we switch vendors?
- What data formats and standards are used for portability?
- What is the contractual notice period and exit process?

Glossary of Key Terms

A reference guide to the technical and regulatory terms used throughout this document.

Term	Definition
Agentic AI	AI systems capable of taking autonomous actions (e.g. booking appointments, updating records) rather than simply generating text responses.
API (Application Programming Interface)	A set of protocols that allows different software systems to communicate and exchange data with each other.
Confidence Threshold	A score (typically 0–1) representing how certain the AI is about its response. Below a set threshold, the system may escalate to a human agent.
CRM (Customer Relationship Management)	Software used to manage customer interactions, records, and data (e.g. Salesforce, Microsoft Dynamics).
Deterministic	A process that produces the same output every time given the same input. Traditional IVR menus are deterministic , pressing 1 always routes to the same place.
EU AI Act	European Union regulation (entered into force 2024) that classifies AI systems by risk level and imposes obligations around transparency, conformity assessment, and prohibited practices.
Fine-Tuning	The process of further training a pre-trained LLM on a specific dataset to specialise its behaviour for a particular domain or task.
Foundation Model	A large AI model (e.g. GPT-4, Claude, Llama) trained on broad data that serves as a base for various downstream applications.
GDPR	General Data Protection Regulation. EU law governing the collection, processing, and storage of personal data. Grants individuals rights including data access, rectification, and erasure.
Guardrails	Rules, filters, or constraints applied to an AI system to prevent it from producing harmful, off-topic, or incorrect outputs.
Hallucination	When an AI model generates a response that sounds plausible but is factually incorrect or entirely fabricated. A key risk in customer-facing applications.
Human-in-the-Loop (HITL)	A design pattern where a human reviews, approves, or intervenes in AI-driven processes , particularly for high-stakes decisions.
ID&V / IDVA	Identity and Verification (sometimes Identity, Document and Verification Assessment). The process of confirming a customer's identity before granting access to their records.
IVR (Interactive Voice Response)	Traditional phone-based system using pre-recorded prompts and keypad/voice inputs to route callers. Deterministic by nature.

LLM (Large Language Model)	An AI model trained on large volumes of text data that can generate, summarise, and interpret natural language. Examples: GPT-4, Claude, Gemini.
Model Drift	Gradual degradation in an AI model's performance over time, often caused by changes in the data it encounters versus the data it was trained on.
Multi-Tenancy	An architecture where a single instance of software serves multiple customers (tenants), sharing infrastructure but logically isolating data.
PII (Personally Identifiable Information)	Any data that could identify a specific individual , names, addresses, phone numbers, account numbers, etc.
Probabilistic	A process whose output can vary even with the same input, due to randomness in the model. AI language models are inherently probabilistic.
Prompt Engineering	The practice of crafting inputs (prompts) to guide an AI model toward desired outputs. Also refers to adversarial manipulation (see Prompt Injection).
Prompt Injection	A security attack where a user crafts input designed to override the AI's instructions, potentially bypassing safety controls or accessing unauthorised data.
RAG (Retrieval-Augmented Generation)	A technique where the AI retrieves relevant information from a knowledge base before generating a response, improving accuracy and grounding.
Red-Teaming	Structured adversarial testing where testers deliberately try to make an AI system fail, produce harmful outputs, or bypass its safety controls.
Responsible AI	A framework of principles and practices ensuring AI is developed and deployed ethically, fairly, transparently, and with appropriate accountability.
RPO (Recovery Point Objective)	The maximum acceptable amount of data loss measured in time. E.g. an RPO of 1 hour means you could lose up to 1 hour of data in a disaster.
RTO (Recovery Time Objective)	The maximum acceptable downtime before a system must be restored after a failure.
SLA (Service Level Agreement)	A contractual commitment defining performance metrics such as uptime, response time, and accuracy targets.
Temperature	A parameter controlling the randomness of an AI model's outputs. Lower temperature = more predictable/conservative; higher temperature = more creative/variable.
Token	The basic unit of text that an LLM processes. A token is roughly ¾ of a word. Pricing and usage limits are often measured in tokens.
Walled Garden	A closed system where data and interactions are contained within a single vendor's ecosystem and not shared externally.