QCS · TECH

WHITEPAPER
29 OCTOBER 2026

AI ENGINEERING

From Demo to Production: Safeguards for Deploying AI Inside Real Apps

Most AI features work in the demo and stall before production. The gap is rarely the model. It is everything around it: who can see what, what happens when the model is wrong, what it costs at scale, and who answers for it when something goes wrong.

7	6	3	18
conditions that change between demo and production	stages in the AI control layer	groups of safeguards: security, quality, operations	checks on the launch readiness list

01 · EXECUTIVE SUMMARY

The gap is rarely the model

This paper sets out the safeguards we recommend, and build, when AI goes inside a business application: an ERP, a project-management suite, a customer portal or a mobile app.

They fall into three groups. Together they turn an impressive demo into a system that a security reviewer, a compliance officer and a business owner can each sign off.

GROUP I

Security and data

The assistant never sees, says or does more than the user it serves could.

GROUP II

Quality and reliability

Answers are grounded in your data, measured before release, and fail safely.

GROUP III

Operations and cost

Every call is logged, budgeted and recoverable when a model or vendor changes.

The paper closes with the Indian regulatory picture as of October 2026, practical lessons from building AI into apps, and a go/no-go checklist for launch.

IN THIS PAPER

02 The demo-to-production gap
Seven conditions that change between demo and production, and what each breaks.

03 Reference architecture
A six-stage control layer between the application and the model.

04 Security and data
Six controls, each mapped to the OWASP Top 10 for LLM Applications.

05 Quality and reliability
Grounding, evaluation sets and human oversight matched to impact.

06 Operations and cost
Logging, budgets, model pinning, quality monitoring and incident plans.

07 The India regulatory lens
DPDP, CERT-In, the 2026 IT Rules, RBI FREE-AI and the EU AI Act.

08 Lessons from the field
Eight lessons from building AI into real applications.

09 Launch readiness checklist
Eighteen go/no-go checks across four areas.

02 · THE DEMO-TO-PRODUCTION GAP

A demo proves the model can answer. Production proves the system can be trusted.

Seven conditions change between the two, and each one breaks something different.

IN THE DEMO	IN PRODUCTION	WHAT GOES WRONG
Hand-picked questions	**Thousands of unscripted questions, typos, Hindi-English mixed in one sentence**	▲ Confident wrong answers on the long tail
One admin account with full access	**Hundreds of users with different roles**	▲ A junior user is shown data meant for finance or HR
Clean sample documents	**Stale, duplicated and contradictory records**	▲ Answers quote last year's policy as current
Read-only answers	**Requests to create, approve and update records**	▲ The wrong record is changed, with no undo
A few calls a day	**Peak loads, retries, long documents**	▲ Cost spikes, timeouts, rate-limit errors
Trusted inputs	**Emails, uploads and web pages carrying hidden instructions**	▲ Prompt injection: the data steers the assistant
One fixed model version	**Providers update and retire models**	▲ Behaviour shifts silently after an upgrade

None of these show up in a two-week proof of concept. All of them show up in the first month of real use.

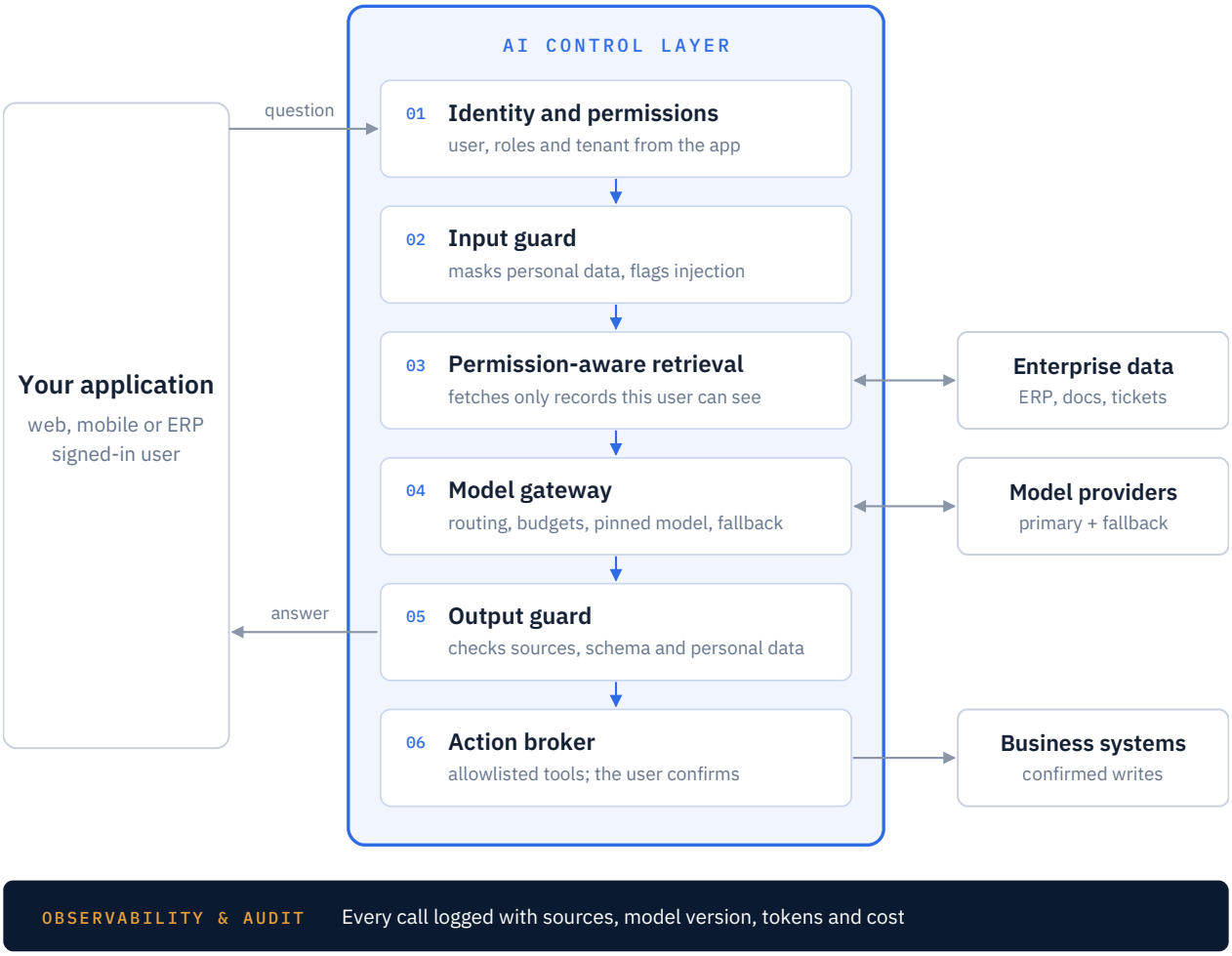
03 · REFERENCE ARCHITECTURE

The model never talks to your users or your data directly

Put a control layer between the application and the model. Every safeguard in the next three sections lives there, so it can be tested, logged and changed without touching the app.

FIGURE 1

Reference architecture for an AI assistant inside a business application



A request passes down the layer in order. The model sees only what the input guard and permission-aware retrieval let through, and nothing reaches the user or your systems without passing the output guard or the action broker.

04 · SAFEGUARDS I

Security and data

The rule is simple: the assistant never sees, says or does more than the signed-in user could do in the application itself.

Six controls enforce it, each mapped to a risk in the OWASP Top 10 for LLM Applications 2025.

S1 · LLM02 · SENSITIVE INFO DISCLOSURE

Permission-aware retrieval

Filter every search by the user's identity and roles from the source system, at query time. Never connect the assistant through a shared service account with full access. Key any cache by user, not by question.

S2 · LLM01 · PROMPT INJECTION

Treat all retrieved content as untrusted

Emails, PDFs, ticket comments and web pages can carry instructions aimed at the model. Keep system instructions separate from data, flag instruction-like text in documents, and never let retrieved text trigger an action on its own.

S3 · DATA MINIMISATION

Minimise personal data before it leaves your boundary

Mask what the task does not need: phone numbers, PAN, bank details, health fields. Where the answer needs the value, tokenise it on the way out and restore it on the way back.

S4 · VENDOR CONTROLS

Configure the model provider like any other data processor

Get contractual no-training and retention terms, pin the hosting region to one you can name to an auditor, use private networking, and keep API keys in a vault with rotation.

S5 · LLM06 · EXCESSIVE AGENCY

Least privilege for actions

Start read-only. Expose write actions only as narrow, allowlisted tools with server-side validation. Require explicit user confirmation for anything that moves money, changes master data or sends a message outside the company.

S6 · LLM05 · IMPROPER OUTPUT HANDLING

Validate output before it touches another system

Treat model output as untrusted user input: escape it for HTML, parameterise any query, and check structured output against a schema.

One more: assume the system prompt will leak (LLM07). Access rules and secrets belong in code, never in the prompt.

05 · SAFEGUARDS II

Quality and reliability

A wrong answer delivered confidently does more damage than no answer. These controls make quality something you measure, not something you hope for.

01 **Ground every answer in your data and show the source.** Retrieve from your own records and cite the document or record behind each claim. When nothing relevant is found, the assistant should say so, not fill the gap. Test for "I don't know" as deliberately as you test for right answers.

02 **Build the evaluation set before the feature.** Collect 100 to 300 real questions with agreed answers, including edge cases, mixed-language queries and deliberate attempts to break it. Run the set on every change to the prompt, model or data, and set a pass mark that gates each release.

03 **Match human oversight to the stakes.** Tier each use case by its impact, then apply the control for that tier (table below).

04 **Fail gracefully.** Set timeouts, qualify a fallback model, and fall back to plain search results or a human handoff. Tell users when the AI is unavailable rather than showing a spinner.

05 **Make the AI visible.** Label AI-generated content, put feedback buttons on every answer, and route flagged answers to a review queue that someone owns.

Human oversight by impact tier

IMPACT TIER	EXAMPLE REQUEST	REQUIRED CONTROL
INFORM	*"Summarise the open RFIs on this project"*	Sources shown; user feedback captured
ADVISE	*"Which vendors are late on deliveries?"*	Sources shown; the user makes the call
ACT	*"Raise a change order for ₹2.4 lakh"*	Preview, explicit confirmation, audit trail, undo where possible
DECIDE ABOUT PEOPLE	Credit, hiring, claims, medical triage	**AI assists only;** a named person decides; reasoning recorded

06 · SAFEGUARDS III

Operations and cost

An AI feature is a live system that keeps changing after launch: users, data and the model itself. Five controls keep it observable, affordable and recoverable.

01 **Log every interaction end to end.** Record the user, prompt version, retrieved document IDs, model and version, tools called, output, latency, tokens and cost. Mask personal data in the logs. Indian rules set the floor: CERT-In requires ICT logs for a rolling 180 days with a copy in India, and the DPDP Rules require processing logs to be kept for at least one year.

02 **Budget per user, per tenant and per feature** (OWASP LLM10, Unbounded Consumption). Set hard caps and alerts, limit context size, cache repeat answers, and route simple tasks to smaller, cheaper models.

03 **Version everything and pin the model.** Keep prompts, model choice and retrieval settings in version control. Re-run the evaluation set before any model switch, and keep a second model qualified so a price change or retirement is a configuration change, not a rebuild.

04 **Watch quality in production, not just uptime.** Review a sample of real conversations every week. Track the share of answers with sources, the refusal rate, thumbs-down rate and human-handoff rate, and alert when any of them drifts.

05 **Plan for an AI incident before you have one.** Define what counts: data shown to the wrong user, a harmful answer, a wrong action taken. Give every AI feature a kill switch, and write the playbook. Specified cyber incidents must reach CERT-In within 6 hours; from May 2027, personal data breaches must reach the Data Protection Board without delay, with a detailed report within 72 hours.

180 days	1 year	6 hours	72 hours
CERT-In: rolling ICT log retention, with a copy in India	DPDP Rules: minimum retention for processing logs	CERT-In: deadline to report specified cyber incidents	DPDP: detailed breach report to the Board, from May 2027

07 · THE INDIA REGULATORY LENS

No AI-specific law, but already regulated

India has no AI-specific law, but an AI feature that touches personal data is already regulated.

PLAN AROUND
13 May 2027

The date when **most DPDP Rules obligations apply**. Consent-manager rules apply earlier, from 13 November 2026.

INSTRUMENT	WHAT IT MEANS FOR AN AI FEATURE	STATUS AS OF OCTOBER 2026
DPDP Act 2023 and DPDP Rules 2025	Notice and consent for personal data the AI uses; reasonable security safeguards; breach intimation to users and the Board; one-year log retention; erasure	Rules notified 13 Nov 2025. Consent-manager rules from 13 Nov 2026. Most obligations from 13 May 2027; a proposal to shorten this to 12 months has not been notified
CERT-In Directions 2022	Report specified cyber incidents within 6 hours; keep ICT logs for 180 days with a copy in India; sync clocks to NIC or NPL time servers	In force since 2022
IT Intermediary Amendment Rules 2026	Platforms that let users create or share AI-generated audio or video must label it, embed permanent metadata, and act on certain orders within 3 hours	Notified 10 Feb 2026; in effect from 20 Feb 2026
India AI Governance Guidelines (MeitY)	Seven principles, from trust and people-first to safety and resilience; relies on existing laws and sector regulators rather than a new AI law	Released November 2025; voluntary
RBI FREE-AI framework	For banks, NBFCs and fintechs: 7 sutras and 26 recommendations covering governance, audit and incident reporting	Released August 2025
EU AI Act (for products serving EU users)	Tell users they are dealing with AI and mark generated content (Article 50); high-risk obligations deferred	Article 50 since 2 Aug 2026; Annex III high-risk from 2 Dec 2027 under Regulation (EU) 2026/1744

For a voluntary baseline, the OWASP Top 10 for LLM Applications, ISO/IEC 42001 (AI management systems) and the NIST AI Risk Management Framework give auditors a shared vocabulary.

This section summarises the rules; it is not legal advice.

08 · LESSONS FROM THE FIELD

The hardest problems are rarely AI problems

They are access, data and integration problems that AI makes visible.

01

Your permissions model becomes your AI's permissions model.

Stale roles and over-broad admin rights were harmless when data was hard to find. An assistant makes it easy to find. Clean up role hygiene before launch, not after the first escalation.

02

Data quality is the ceiling.

The assistant will quote an outdated SOP as confidently as the current one. Archive old versions, tag documents with an owner and a review date, and show "last updated" next to every citation.

03

Start read-only, then earn write access.

Ship search, Q&A and summaries first. Add actions one at a time, each with its own evaluation cases and confirmation step.

04

Integration is most of the work.

Connecting to ERP and legacy APIs, mapping identities, and handling rate limits take far longer than the prompt. Plan the project around the plumbing.

05

Users test the boundaries in week one.

Expect someone to ask for salaries, a colleague's records or the system prompt. Red-team those questions before launch and keep them in the evaluation set.

06

Enterprise security reviews ask the same questions every time.

Where does data go, which region, who can see the logs, is it used for training, how long is it kept. Keep an AI security pack ready: data-flow diagram, provider terms, retention policy, test results.

07

Cost surprises come from context, not users.

Long documents and growing chat history drive most token spend. Summarise history, cap context and measure cost per conversation from day one.

08

Indian users mix languages.

Hindi-English queries and regional terms are normal. Put them in the evaluation set, or the assistant will fail the users who need it most.

09 · LAUNCH READINESS CHECKLIST

Go / no-go

An AI feature is ready to ship when every box below is ticked. Any unticked box is a launch blocker or a written, owned exception.

SECURITY AND DATA 5

- [ ] Retrieval enforces the signed-in user's permissions, tested with at least three roles
- [ ] Prompt-injection tests pass on documents, emails and uploads
- [ ] Personal data not needed for the task is masked before it reaches the model
- [ ] Provider contract confirms no training on your data, retention terms and hosting region
- [ ] Write actions are allowlisted, validated server-side and confirmed by the user

QUALITY 4

- [ ] Evaluation set of real questions passes the agreed mark on the release build
- [ ] Every use case is assigned an impact tier with its control in place
- [ ] The assistant says "I don't know" when sources are missing
- [ ] Fallback path works when the model is slow or down

OPERATIONS 5

- [ ] Interactions are logged with model version, sources, tokens and cost
- [ ] Budgets and alerts are set per user, tenant and feature
- [ ] Model version is pinned and a second model is qualified
- [ ] A named owner reviews flagged answers every week
- [ ] Kill switch tested; AI incident playbook written

COMPLIANCE 4

- [ ] Privacy notice covers AI processing of personal data
- [ ] Log retention meets CERT-In (180 days) and DPDP (one year) requirements
- [ ] Users are told when they are dealing with AI and AI content is labelled
- [ ] Sector rules checked (RBI, SEBI, IRDAI, health) where they apply

SOURCES

1. MeitY notifies final DPDP Rules 2025, S.S. Rana & Co.
2. DPDP Rules, Rule 7: breach intimation
3. Is the DPDP 18-month window being cut to 12 months?, reviewed October 2026
4. CERT-In directions for Indian enterprises, ManageEngine
5. Rules on AI-generated content, labels and faster takedowns, Business Today
6. India AI Governance Guidelines 2025, MediaNama
7. RBI FREE-AI committee report, KPMG India
8. AI Act obligations and delays under Omnibus 2026/1744, ActuIA
9. OWASP Top 10 risks for LLM security 2025, Invicti

ABOUT QCS

Applications built end to end, with AI designed in from the first sprint.

QCS Tech Private Limited builds applications end to end, from web and mobile apps to enterprise integrations, and embeds AI into them with the safeguards in this paper designed in from the first sprint. A typical AI engagement runs in four stages.

STAGE 01

Readiness assessment

Use cases, data sources, permissions and impact tier for each.

STAGE 02

Read-only pilot

Grounded Q&A over real data, measured against an evaluation set.

STAGE 03

Production hardening

The control layer, logging, budgets and a security pack for your reviewers.

STAGE 04

Run and improve

Quality monitoring, model updates and new actions added one at a time.

To discuss an AI deployment inside your application, contact the QCS team.

QCS · TECH