

The Go-Live Gate.

Forty questions to answer before an AI pilot goes into production. Twelve of them are gates: if any one is still a no, the pilot isn't ready, however good the demo looked.

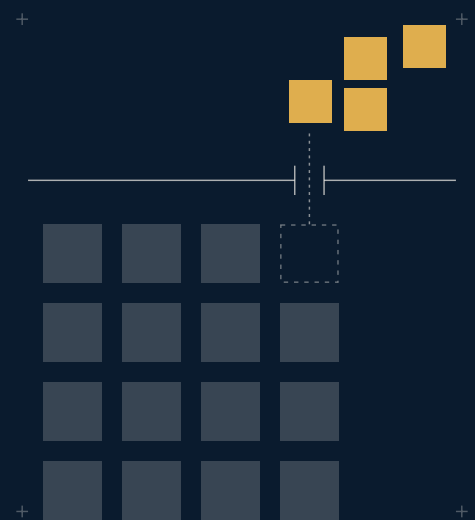


Fig. 01 · One pilot through the gate.
The rest wait for evidence.

Why pilots stall at the last step.

Nobody's AI programme fails because of the model. Pilots stall on questions nobody answered while the demo was going well.

i The pilot was never built to ship Demos are tuned for the happy path. Production isn't. Without an evaluation harness nobody can say whether the system is good enough, so nobody signs it off, and it stays a prototype indefinitely.	ii There is no number for the board "It feels faster" does not survive a budget review. If value wasn't measured from the start, the programme has nothing to defend itself with at the moment it needs funding to scale.	iii The approach was sized for a different company Enterprise frameworks are overkill for a 300-person business. Startup-speed delivery is reckless inside a regulated one. Most advice is calibrated for an organisation that isn't yours.
---	--	--

This checklist turns those failures into 40 questions across eight areas. **12 are gates**, marked with an amber-edged box: they must all be true before real users or real decisions depend on the system. The other 28 decide whether you ship cleanly or ship with conditions.

How to run the review

- 1 Four to six people, 60 minutes.** The business owner, the engineering lead who built the pilot, whoever will run it in production, and someone from risk or security. If the person who gets called when it breaks isn't in the room, your answers on operations are guesses.
- 2 About a minute per item.** Record the answer, the owner and the evidence. Don't fix anything in the meeting; save the time for arguing about gates.
- 3 Tick only what you could evidence today.** Each item is a statement your team should be able to make truthfully. "We think so" counts as no. So do "nearly", "it's in the backlog" and "the vendor says it does".
- 4 Leave with a name against every open item.** The owners shown are suggested roles. In a 300-person firm one person may cover three of them; a regulated enterprise may need a formal artefact for each gate. The questions stay the same.

Plenty of people can build an impressive prototype. Far fewer can get one through a procurement cycle, a compliance function and a sceptical engineering team, and still have it running six months later.

EDDIE FORSON · FOUNDER, KISEKI LABS

40 questions, 12 gates.

#	AREA	ITEMS	GATES	THE QUESTION BEHIND IT
01	Value and ownership	5	2	Who owns this, and how will we know it worked?
02	Evaluation and ground truth	5	2	How do we know the output is correct?
03	Data and security boundaries	5	1	Where does our data go, and what can the system touch?
04	Failure behaviour and human review	5	2	What happens when it's wrong?
05	Cost and performance	5	1	What does each transaction cost, and is it fast enough?
06	Operations and change	5	2	How will we know when it degrades, and what do we do then?
07	Handover and ownership	5	1	What happens when the builders leave?
08	Governance and compliance	5	1	Who signs it off, and what is on the record?
Total		40	12	

☐ ITEM ☒ GATE: MUST BE TRUE BEFORE GO-LIVE

01 · 5 ITEMS · 2 GATES

Value and ownership

Who owns this, and how will we know it worked?

If nobody owns the outcome and nobody wrote down what success looks like, the go-live decision becomes a matter of opinion. In practice that means the opinion of the most senior person in the room.

☒ 1.1 One named person owns the business outcome and has the authority to stop the project.

A pilot without an owner drifts. Nobody signs it off, and nobody switches it off either.

GATE · OWNER: EXECUTIVE SPONSOR

☒ 1.2 The success metric, its target and a review date were written down before the results came in.

Pre-registering the bet stops a disappointing result being redefined as a success after the fact.

GATE · OWNER: BUSINESS OWNER

☐ 1.3 There is a measured baseline for the current process: time, cost and error rate on real work.

"It feels faster" does not survive a budget review. The improvement is only as credible as the baseline.

OWNER: OPERATIONS LEAD

☐ 1.4 The business case counts the outcome of the whole workflow, not the speed-up on a single task.

A step that runs ten times faster rarely makes the job ten times faster. The board is paying for the job.

OWNER: FINANCE PARTNER

☐ 1.5 There is a written kill criterion: the result, or the date, at which the project stops or is rethought.

Projects that cannot fail never get stopped. They just stop being mentioned.

OWNER: EXECUTIVE SPONSOR

Evaluation and ground truth

How do we know the output is correct?

This is the section most pilots fail, and the one that makes every other argument cheaper. Accuracy is stated against a named test set, or it isn't stated.

- ☒ **2.1 There is a fixed test set of real cases, each with its expected answer, agreed with the people who do the work today.**

This is what "correct" means, written down. Without it, accuracy is an opinion and every debate is settled by seniority.

GATE • OWNER: PRODUCT OWNER, WITH DOMAIN EXPERTS

- ☒ **2.2 The pass threshold was agreed by the business owner before the final run, and the version going live meets it, overall and for each type of case.**

A bar set after you have seen the score isn't a bar. A healthy average can also hide one category that fails badly.

GATE • OWNER: BUSINESS OWNER

- ☐ **2.3 The test set covers edge cases, unhappy paths and inputs the system should refuse, as well as the everyday questions.**

Demos are tuned for the happy path. Production isn't, and neither are your users.

OWNER: ML / AI LEAD

- ☐ **2.4 The evaluation runs automatically in CI on every change to prompts, models, code or configuration, and a drop in score blocks the release.**

A prompt tweak that fixes one case can quietly break five others. Until this is automated, run the suite by hand before every release.

OWNER: ENGINEERING LEAD

- ☐ **2.5 Where a model grades outputs, its grades have been checked against human judgement on a sample.**

An automated judge is a model too. If nobody has checked it, you are measuring one guess with another.

OWNER: ML / AI LEAD

Data and security boundaries

Where does our data go, and what can the system touch?

Data boundaries are a design input. Settled at architecture time, they cost a page of documentation. Discovered in the security review the week before launch, they cost the launch date.

- ☒ **3.1 There is a written data-flow map showing which providers and services see which data, and what never leaves your perimeter.**

Your security review will ask for it. Answering at design time is cheap; answering after launch means a retrofit.

GATE • OWNER: SECURITY LEAD

- ☐ **3.2 Each provider's retention, training-use and processing-region terms are confirmed in the contract and acceptable for the data involved.**

Default terms differ by provider, tier and product, and they change. A marketing page is not a contract.

OWNER: DATA PROTECTION LEAD, WITH PROCUREMENT

- ☐ **3.3 Credentials live in a secrets manager, are scoped to this system, and have been rotated at least once.**

A leaked key is both a bill and a data incident. Rotation is only quick if someone has done it before.

OWNER: PLATFORM LEAD

- ☐ 3.4 **Every point where outside content (uploads, documents, emails, web pages) meets something that can execute or act has been listed and contained.**

Data pipelines are an attack surface: a file that arrives as data can run as code, or instruct an agent with tools.

OWNER: SECURITY LEAD

- ☐ 3.5 **The system's permissions are the minimum its job needs, enforced by the platform rather than described in a prompt or a policy.**

Agents act on the controls that exist, whatever the policy document says. A permission the system doesn't have is the control that holds.

OWNER: SECURITY LEAD

04 · 5 ITEMS · 2 GATES

Failure behaviour and human review

What happens when it's wrong?

It will be wrong sometimes. The aim is a system that fails visibly instead of silently, with a person in the loop wherever a mistake is expensive.

- ☒ 4.1 **We have defined what the system does when it is unsure or fails (refuse, retry, escalate or fall back), and every failure is logged and visible rather than passed on as an answer.**

A silent wrong answer costs more than an honest "I can't answer that". Failure behaviour is either designed in advance or discovered by a customer.

GATE · OWNER: PRODUCT OWNER

- ☐ 4.2 **Outputs pass a verification layer before anything acts on them: schema checks, business rules, or a second check by a model.**

Deterministic checks are cheap, fast and never hallucinate. Use them before you reach for anything cleverer.

OWNER: ENGINEERING LEAD

- ☒ 4.3 **A person reviews the output wherever being wrong is expensive, has the time and authority to overrule it, and the system cannot skip that step.**

Human review by design is often what makes a system deployable in a regulated process. A rubber stamp is not a control.

GATE · OWNER: RISK AND COMPLIANCE

- ☐ 4.4 **When the system escalates, it names the specific discrepancy or reason, so the reviewer has one thing to check.**

Reviewers told what doesn't reconcile resolve problems. Reviewers told only that something is wrong go hunting.

OWNER: PRODUCT OWNER

- ☐ 4.5 **Users can flag a wrong answer in one step, and someone reads the flags every week.**

Users see failures first. Make reporting one cheaper than working around it, and every flag becomes a test case.

OWNER: PRODUCT OWNER

Cost and performance

What does each transaction cost, and is it fast enough?

Pilot economics are dangerous to extrapolate. Inputs get longer, retries happen and volume changes shape, so measure the unit cost on realistic work before anyone commits to a budget.

- ☒ **5.1 The cost per transaction (per query, document or case) is measured on realistic runs, including retries, and the business case holds at that cost and at expected volume.**

The demo's unit cost is not a forecast. A business case built on it is a guess that finance will eventually check.

GATE · OWNER: FINANCE PARTNER

- ☐ **5.2 Spend is attributed to this system, with a monthly budget and alerts at agreed thresholds.**

You cannot govern spend you cannot see. Alerts start a conversation; a surprise invoice starts an inquiry.

OWNER: FINANCE PARTNER, WITH PLATFORM LEAD

- ☐ **5.3 Each step runs on the smallest model that clears its quality bar on your test set.**

The most capable model is often the slowest and most expensive way to do a simple step.

OWNER: ML / AI LEAD

- ☐ **5.4 There is an agreed response-time target, and it is met at realistic load with retries included.**

Users route around slow tools. Retries that rescue accuracy also add seconds, so measure the two together.

OWNER: ENGINEERING LEAD

- ☐ **5.5 The system has been load-tested at peak volume, and the provider's rate limits cover it.**

A rate limit is a capacity ceiling someone else controls. Find it in testing rather than on launch day.

OWNER: PLATFORM LEAD

Operations and change

How will we know when it degrades, and what do we do then?

Models change, inputs drift and providers have bad days. A production system has to show you when its quality moves, and give you somewhere to go when its model does.

- ☒ **6.1 Production quality is monitored as well as uptime: a sample of live outputs is scored regularly against the same standard as the test set, with an alert when it drops.**

A system can be up and wrong for weeks. An uptime dashboard will not tell you.

GATE · OWNER: OPERATIONS LEAD

- ☐ **6.2 Model versions are pinned, and any upgrade goes through the evaluation like any other change.**

A model that updates silently is an untested release you didn't schedule.

OWNER: ENGINEERING LEAD

- ☒ **6.3 There is a failover plan for when the model is withdrawn or the provider is unavailable: a fallback measured on your test set, or a manual process the business has agreed to run.**

Models get retired and restricted, sometimes at a few hours' notice. An unmeasured fallback is a guess.

GATE · OWNER: ENGINEERING LEAD

- ☐ **6.4 Prompts, model settings and configuration are version-controlled with the code, and the whole set can be rolled back in minutes. Someone has done it.**

A prompt is code that changes behaviour. A rollback nobody has run is a theory.

OWNER: PLATFORM LEAD

- ☐ **6.5 Real failures, user flags and reviewer overrides are added to the test set, so it gets harder over time.**

A test set frozen at launch measures the problems you had in the pilot, not the ones you have now.

OWNER: ML / AI LEAD

07 · 5 ITEMS · 1 GATE

Handover and ownership

What happens when the builders leave?

Whoever built the pilot may not be around in six months. Plan for that before go-live, while it is cheap.

- ☒ **7.1 A named team owns the system in production, with on-call cover, a budget and the skills to change it.**

Ownership means a team with a rota. A person who happens to remember how it works is a risk.

GATE · OWNER: HEAD OF ENGINEERING

- ☐ **7.2 Documentation is written for the engineers who will inherit the system: architecture, design decisions and the reasons for them, configuration and known limits.**

Decisions recorded without their reasons get reversed by the next person, usually at some cost.

OWNER: ENGINEERING LEAD

- ☐ **7.3 The tests and evaluation suite live in the owning team's repository, and that team runs them unaided.**

If only the builder can run the evaluation, the evaluation leaves when the builder does.

OWNER: ENGINEERING LEAD

- ☐ **7.4 There is a runbook covering what to check, who to call, and how to switch the system off safely.**

The first incident is the wrong moment to work out where the off switch is.

OWNER: OPERATIONS LEAD

- ☐ **7.5 Knowledge transfer is scheduled, and at least two people in the owning team have shipped a real change to the system.**

Shipping a change teaches what documents can't, and two people means no single point of failure.

OWNER: HEAD OF ENGINEERING

08 · 5 ITEMS · 1 GATE

Governance and compliance

Who signs it off, and what is on the record?

Risk and compliance ask the right questions, usually late. Involve them early and sign-off becomes routine.

- ☒ **8.1 Risk, legal and security have reviewed the system, and the sign-off path was agreed with them early.**

Their questions are the right ones. Asked in the final week, they turn into a delay nobody budgeted for.

GATE · OWNER: RISK AND COMPLIANCE

- ☐ **8.2 Key decisions are recorded with their reasons: model choice, data boundaries, thresholds, and what a person reviews.**
When someone asks why the system behaves as it does, a decision log answers in minutes.
OWNER: PRODUCT OWNER
-
- ☐ **8.3 The system is in your AI register or inventory, with its owner, purpose, data, risk rating and next review date.**
You cannot govern AI you cannot list. If there is no register yet, this system is a good place to start one.
OWNER: RISK AND COMPLIANCE
-
- ☐ **8.4 People affected by the system's outputs are told AI is involved where law, contract or your own policy requires it, and know how to challenge a decision.**
Some automated decisions carry legal duties to inform people and offer human review. Ask legal which apply.
OWNER: LEGAL / DATA PROTECTION LEAD
-
- ☐ **8.5 There is a route for reporting incidents and near misses, including the times a person caught a bad output before it did harm, and someone owns acting on them.**
Near misses are the cheapest lessons you will get. Without a reporting route, you only learn from the incidents.
OWNER: RISK AND COMPLIANCE
-

WORKED EXAMPLE · KISEKI LABS ENGAGEMENT

From under 60% to 98% in four weeks.

A data analytics start-up's platform for plain-English questions about spreadsheet data completed fewer than 60% of queries after months of development. Here is how the four weeks map to this checklist.

98% Query success on the fixed suite, up from under 60%.	35 Test cases across six business domains in the evaluation suite.	79 Passing tests handed over: 58 unit and 21 integration.	0 Syntax errors reaching execution, thanks to a validation gate.
WEEK 1 · FOUNDATIONS FIRST Dependency management, linting, CI and documentation, before touching the architecture. Everything after was easier to test and deploy. 2.4	WEEK 2 · GROUND TRUTH A fixed 35-case suite across six business domains settled the main architecture question on evidence. 2.1 · 2.2	WEEK 3 · BUILT TO FAIL VISIBLY Four agents, a validation gate and self-healing retries. Typical responses fell from over 60 seconds to around 25, helped by a faster planning model. 4.1 · 4.2 · 5.3	WEEK 4 · HANDED OVER Tests, the evaluation suite, documentation and knowledge transfer to the engineers who would own the system. 7.2 · 7.3 · 7.5

Not every box was ticked in four weeks. The handover listed nine prioritised areas for future work, with code-execution sandboxing flagged as security-critical (item 3.4). In this checklist's terms, that is a condition with an owner and a date, which is worth far more than a box ticked generously. [Read the full write-up.](#)

Score it honestly, then decide.

Count two numbers: gates passed out of 12, and items passed out of 40, gates included. The gates decide whether you can ship. The total decides how.

- **A gate passes only if it is fully true today.** Gates can't be marked N/A: each is written to apply to any system that carries a real business process.
- **An ordinary item can be N/A** when it genuinely cannot apply. Write the reason down and take it out of the total: if four items don't apply, score out of 36, and the thresholds become 29 and 22.
- **Don't average your way past a gate.** 39 out of 40 with the test set missing is still Not yet.

VERDICT	WHEN	WHAT TO DO
Ship	All 12 gates, and 32 or more of 40 items (80% or more).	Put every open item on a dated list with an owner. Book the first production review for 30 days after launch and score the checklist again there.
Ship with conditions	All 12 gates, and 24 to 31 of 40 items (60% to 79%).	Write each condition down with an owner and a date, ideally within 60 days. Limit exposure until they close: a named user group, capped volume, or running alongside the current process. If a condition misses its date, the verdict reverts to Not yet.
Not yet	Any gate open, or fewer than 24 of 40 items.	Close the gates first. They are usually weeks of work rather than months. If the pressure to launch is real, run the system in shadow, alongside the current process with nobody acting on its output, while you close them. Don't call that go-live.

WHAT WE SEE MOST OFTEN

The gate we find missing most often is the fixed test set (2.1). It is also the one that makes every other argument cheaper: once correct is defined and measured, debates about models, prompts and architecture get settled by evidence rather than seniority.

A note on the floor. Passing all 12 gates with most of the other items open is unusual, because the items are what make the gates true. If that is your result, check whether someone answered the gates generously.

A tracker spreadsheet comes with this checklist. It holds the same 40 items with owners, status, due dates and evidence links, and works out the verdict as you go.

Go-Live Gate summary sheet

SYSTEM OR PILOT

REVIEW DATE

BUSINESS OWNER

CHAIRER BY

#	GATE	PASSED	OWNER	EVIDENCE OR NOTE
1.1	Named owner who can stop the project	☐ Yes ☐ No		
1.2	Success metric and target set in advance	☐ Yes ☐ No		
2.1	Fixed test set of real cases with expected answers	☐ Yes ☐ No		
2.2	Agreed pass threshold met, overall and by type	☐ Yes ☐ No		
3.1	Data-flow map: who sees what, what stays inside	☐ Yes ☐ No		
4.1	Failure behaviour defined; every failure visible	☐ Yes ☐ No		
4.3	Human review wherever being wrong is expensive	☐ Yes ☐ No		
5.1	Cost per transaction measured; business case holds	☐ Yes ☐ No		
6.1	Production quality monitored, with alerts	☐ Yes ☐ No		
6.3	Failover: measured fallback or agreed manual process	☐ Yes ☐ No		
7.1	Named owning team with on-call cover	☐ Yes ☐ No		
8.1	Risk, legal and security sign-off path agreed	☐ Yes ☐ No		

GATES PASSED

ITEMS PASSED (OF 40, LESS ANY N/A)

NEXT REVIEW DATE

____ / 12

____ / ____

 VERDICT ☐ Ship ☐ Ship with conditions ☐ Not yet

CONDITIONS OR OPEN GATES • OWNER • DUE DATE

BUSINESS OWNER

RISK AND COMPLIANCE

ENGINEERING LEAD

NAME

NAME

NAME

SIGNATURE AND DATE

SIGNATURE AND DATE

SIGNATURE AND DATE

If the review turned up open gates.

That is what the review is for. Closing gates is usually weeks of work, and the order matters: ground truth first, then everything that depends on it.

PILOT TO PRODUCTION • 4 TO 8 WEEKS • FIXED FEE

Get the pilot through the gate.

Pilot to Production is our most requested engagement: four to eight weeks, fixed scope and fixed fee. We build the evaluation harness and a measured baseline, take the system to a production architecture with tests, CI and operational documentation, and hand it over to your engineers.

If your review turned up open gates, start with a 30-minute briefing. You'll get a straight answer on whether the work is worth doing, including when the answer is that you don't need us.

[REQUEST AN EXECUTIVE BRIEFING → KISEKILABS.COM/BOOK-A-CALL](https://kisekilabs.com/book-a-call)

EARLIER IN THE JOURNEY?

No pilot yet, or still choosing which one to back? Start with the [AI Readiness Assessment](https://kisekilabs.com/ai-readiness-assessment) at kisekilabs.com/ai-readiness-assessment.

SEE THE METHOD IN PRACTICE

The full write-up of the engagement in the worked example, week by week: [From 60% to 98%: what a 4-week AI consulting engagement actually looks like](#).

ABOUT KISEKI LABS

AI advisory and delivery for mid-market and enterprise organisations in the UK and USA. Founder Eddie Forson has built production AI systems since 2017, from computer vision to LLM agents, much of it inside regulated financial services. Evidence first, architecture second.

Edition 01, October 2026. You are welcome to copy and adapt this checklist for use inside your organisation. kisekilabs.com