

The AI Vendor Evaluation Checklist

How to choose the right partner to build, integrate, and run
AI in your business

A decision tool for CEOs, CTOs, and operations leaders by Tech.us



Why this guide exists

Most AI failures aren't technology failures. They're selection failures.

A company decides it needs AI, sits through four impressive demos, picks the vendor with the slickest deck, and eighteen months later has spent real money on something that never reached production, or worse, reached production and quietly started making bad decisions nobody caught.

The pattern is well documented. MIT's 2025 *State of AI in Business* study found that the large majority of enterprise generative-AI pilots delivered no measurable return, not because the models couldn't perform, but because the work was scoped against the wrong problem, handed to the wrong builder, or deployed without anyone watching it over time.

This guide is built to prevent that. It walks you through how to evaluate the company you're trusting to build or run your AI, including what to ask, what good answers sound like, and what should make you walk away. The workbook at the end turns all of it into a scorecard you can run against any vendor in about an hour.

A note on language: throughout this guide, "**vendor**" means *whoever you're relying on to deliver AI capability*, an AI development partner, a systems integrator, a managed-AI provider, or a platform/model provider. It is not limited to off-the-shelf software. As you'll see in the next section, for most decisions that actually matter, off-the-shelf software is the *least* likely answer.

How to use this document

- Read the guide once to understand what you're evaluating and why.
- Use the red-flag list as a fast first filter to disqualify weak vendors early.
- Use the workbook to score your shortlist (aim for 3–5 candidates, then 2–3 finalists).
- Keep the completed workbook on file. You'll want it when a stakeholder asks why you chose Vendor A over Vendor B.

Before anything else: Build, Buy, or Partner?

Vendor evaluation starts one step earlier than most people think. Before you compare providers, you need to know what kind of relationship you're actually buying. There are three:

Buy an off-the-shelf tool.

You license a finished product and use it as-is. Fast and cheap, but you get exactly what every competitor using the same tool gets. Sensible only for commoditized, non-differentiating tasks such as transcription, generic chat support, document summarization, where "good enough and standard" is genuinely fine.

Build in-house.

You own the outcome completely, but you also own the cost, the hiring, the maintenance, and the risk of falling behind as the technology moves. Realistic only when you already have spare AI engineering capacity and a use case so specific to your business that nothing external fits.

Partner with an AI company to build for you.

You get custom-fit capability and ownership of the result without standing up an internal AI team from scratch, and a partner who stays current on models, security, and compliance so you don't have to. For most mid-market and enterprise organizations, this is the path that produces differentiated AI that actually ships.

Here's the part the generic checklists won't tell you: **the use cases that move your business are almost never the ones you can buy off a shelf.** If a capability is available as a standard product, it's available to your competitors too, which means it can't be a source of advantage.

The moment AI touches something proprietary to you, your data, your workflows, your customer relationships, an agentic process that has to act inside your systems, you've left the off-the-shelf category.

You're now choosing between building and partnering, and the rest of this guide is built for that decision.

	Off-the-shelf (Buy)	Build in-house	Partner to build
Best for	Commodity tasks, no differentiation needed	Truly unique need + existing AI team	Differentiating or agentic use cases
Time to value	Fastest	Slowest	Fast, with real fit
Differentiation	None — competitors get the same thing	High	High
You own the IP / models	No	Yes	Negotiable — and you should insist on it
Ongoing maintenance burden	Vendor	You	Shared / vendor
Biggest risk	Generic results, lock-in	Cost, talent, falling behind	Choosing the wrong partner

If you've concluded you're building or partnering, the rest of this guide tells you how to choose well.

First: Define what success looks like before you talk to anyone

The most expensive mistake in AI isn't picking the wrong vendor. It's picking a capable vendor to solve a problem that wasn't worth solving. That decision gets made before the first meeting, when you're still deciding what you're trying to achieve.

Answer these for yourself first. A good partner will ask you the same questions, and the ones who *don't* are telling you something.

- **Which 2–3 specific workflows or decisions are we targeting?**
Not "adopt AI." A named process with a named owner.
- **What does measurable success look like at 90 days, and at 12 months?**
Tie it to a number like hours saved, error rate, cycle time, revenue, cost.
- **What does our current environment look like?**
Core systems, data sources, where the data lives, and how clean it is.
- **Who internally owns this —**
both the rollout and the long-term operation?
- **What's the real budget,**
including integration, data work, training, and ongoing run cost, not just the build?

If you can't answer these, no vendor can rescue the project. Start here.

The eight dimensions of a strong AI vendor

This is the core of the evaluation. For each dimension you'll find why it matters, the questions to ask, and how to tell a strong answer from a weak one. The workbook turns each into a scored worksheet.

1. Strategic & domain fit

Why it matters: A technically brilliant partner who doesn't understand your business will build the wrong thing beautifully. Fit with your problem, industry, and outcomes predicts success better than raw technical horsepower.

Ask:

- Walk us through how you'd approach *our* problem — before we talk technology.
- Where have you solved something structurally similar? What was the outcome?
- How do you measure whether what you build actually worked?

Strong answer: They lead with your business problem and outcomes, reference comparable work, and talk in terms of results, not features.

Weak answer: They jump straight to their tech stack and pitch capabilities you didn't ask for.

2. Technical capability & architecture

Why it matters: You're hiring engineering judgment, not a feature list. How a partner chooses models, designs the system, and proves it works tells you more than any demo.

Ask:

- How do you decide which model or approach fits a given problem, and how do you avoid lock-in to a single provider?
- How do you design for explainability? Can you show *why* the system produced a given output?
- If this involves agents acting in our systems, how is the architecture scoped and controlled?
- How will this scale as our usage grows 2–3x?

Strong answer: Pragmatic, model-agnostic reasoning; explainability designed in; honest about limitations.

Weak answer: "It's proprietary, we can't explain how it works." Black-box opacity is disqualifying, not impressive.

3. Data governance & IP ownership

Why it matters: This is where build/partner engagements differ most from buying software, and where companies get burned. You need to know where your data goes, and you need to know who owns what gets built.

Ask:

- Where is our data stored and processed, and who can access it?
- Will our data be used to train your models or anyone else's? Is there a written opt-out?
- **Who owns the custom code, models, prompts, and pipelines you build for us?**
- What's the data export and deletion process if we leave?

Strong answer: Clear data handling, no training on your data without consent, and a contract that assigns the custom IP to *you*.

Weak answer: Vague on data usage, or expects to retain ownership of work you paid them to build.

4. Security & compliance

Why it matters: Standard certifications are necessary but no longer sufficient. AI introduces risks, training-data provenance, automated-decision accountability, bias, that traditional security reviews miss entirely.

Ask:

- How do you align with the NIST AI Risk Management Framework, and how do you handle EU AI Act obligations if they apply to us?
- What's your process for bias and fairness testing? Can you produce audit trails for AI-driven decisions?

Strong answer: Can speak specifically to NIST AI RMF and EU AI Act, documents bias testing.

Weak answer: Treats compliance as your problem, or dismisses it as "not relevant yet."

5. Model reliability & drift control

Why it matters: AI is not a build-once asset. Models degrade as the world changes, a phenomenon called drift, and agentic systems can quietly stray from their original goal. The strongest vendors plan for the day the system stops working, not just the day it ships. This single dimension separates a current AI partner from a dated one.

Ask:

- How do you monitor for model drift and accuracy degradation over time, and what triggers a retrain?
- Can you share evidence of **red-team or adversarial testing**? How recent, and what did it find?
- For agentic systems: how do you prevent goal drift, control which tools and integrations an agent can touch, and keep a human in the loop on high-stakes actions?
- Is there a kill-switch and a rollback plan?

Strong answer: Continuous monitoring, scheduled red-teaming with findings they'll discuss, explicit human-in-the-loop and rollback controls.

Weak answer: "We tested it before launch." One-time testing is not a control. New attack methods and drift appear continuously.

6. Delivery, integration & implementation realism

Why it matters: A capable system that doesn't connect to your real systems delivers impressive demos and zero operational value. And timelines are systematically optimistic, independent implementation data consistently shows real deployments running roughly 2–4x longer than the vendor's slide.

Ask:

- Who is actually on our team — names, roles, seniority? Or does the A-team disappear after signing?
- How will this integrate with our existing stack? Have you connected to these systems before?
- Will you run a proof-of-concept with *our* real data before we commit?
- What does knowledge transfer look like, so we're not permanently dependent on you?

Strong answer: A named, credible team; a real POC with your data; a realistic timeline that survives scrutiny.

Weak answer: Promises a complex, multi-workflow build live in a few weeks. That's either a misunderstanding of your business or a willingness to misrepresent it.

7. Commercial & contractual

Why it matters: AI cost models can be volatile, and standard software contracts don't cover AI-specific risk. The contract is where reliability promises either become enforceable or evaporate.

Ask:

- What's the total cost of ownership — build, integration, run cost, retraining — not just the headline price?
- If pricing is usage-based, what caps and alerts protect us from runaway bills?
- **Who is liable when the AI gets something wrong?** Get the accountability split in writing.
- What are the SLAs, the exit terms, and the data-portability guarantees?

Strong answer: Transparent TCO, defined liability, real SLAs, and a clean exit with your data and IP intact.

Weak answer: Evasive on cost, silent on liability, or heavy lock-in with no exit path.

8. References & track record

Why it matters: A reference call with a customer like you is the single most reliable signal you'll get. It's more honest than any case study, because the questions are yours.

Ask the vendor:

- Can we speak with three customers similar to us in industry, size, and use case?

Ask their references:

- What was the real timeline versus what you were quoted?
- How responsive is support when something actually breaks?
- Were there costs you didn't see coming?
- Knowing what you know now, would you choose them again?

Strong answer: Produces relevant production references readily.

Weak answer: Can't name three comparable customers in production. Move on.

Red flags: when to walk away

Any one of these is a reason to stop, regardless of how good the demo looked.

- ☐ **Can't explain how the AI works** in plain language, or hides behind "proprietary."
- ☐ **Won't commit to assigning the custom IP**, including code, models, prompts to you.
- ☐ **No evidence of red-team or adversarial testing**, or treats one-time pre-launch testing as sufficient.
- ☐ **Promises full automation with no human oversight**, or "guaranteed ROI" on a fixed date.
- ☐ **Promises a complex build in production in a few weeks**. Narrow single-purpose tools can deploy fast; multi-workflow or agentic builds cannot.
- ☐ **Dismisses security, compliance, or bias concerns** as not their problem or not relevant.
- ☐ **Vague or evasive on pricing**, total cost, or future cost changes.
- ☐ **Can't produce production references** in your industry or size category.
- ☐ **Silent on liability** — won't define who's accountable when the system errs.
- ☐ **Heavy lock-in, no trial, no clean exit**, or proprietary data formats you can't take with you.

How to score: the weighted scorecard

Score each of the eight dimensions from 1 to 5:

- **1** — Very weak or not demonstrated
- **3** — Adequate, with gaps
- **5** — Strong evidence, fully convincing

Then weigh the dimensions. For a build-or-partner engagement, the dimensions that most predict success carry the most weight. Recommended weighting (adjust to your situation):

#	Dimension	Weight
1	Strategic & domain fit	20%
2	Technical capability & architecture	15%
3	Data governance & IP ownership	15%
4	Security & compliance	10%
5	Model reliability & drift control	15%
6	Delivery, integration & implementation realism	15%
7	Commercial & contractual	5%
8	References & track record	5%
	Total	100%

Multiply each score (1–5) by its weight, then sum. A weighted total above **4.0** is a strong candidate; **3.0–4.0** is workable if the gaps are in lower-weighted dimensions; below **3.0**, or any single **1** in a top-weighted dimension, is a pass.

Shortlist 3–5 vendors, score your top 2–3 in full, and run a 2–4 week proof-of-concept with your real data before you sign anything.

The Workbook

Print this section and complete one set of worksheets per vendor. Budget about an hour per vendor.

Vendor snapshot

- Vendor name: _____
- Primary contact / role: _____
- Engagement type (build / partner / buy): _____
- Use case being evaluated: _____
- Date of evaluation: _____ Evaluator: _____

Worksheet 1: Strategic & domain fit

Led with our problem, not their tech? ☐ Yes ☐ No

Relevant comparable work shown? ☐ Yes ☐ No

Defines success by outcomes, not features? ☐ Yes ☐ No

Notes: _____

Score (1–5): _____

Worksheet 2 : Technical capability & architecture

Model-agnostic / avoids single-provider lock-in? ☐ Yes ☐ No

Can explain *why* the system produces an output? ☐ Yes ☐ No

Honest about limitations? ☐ Yes ☐ No

Scales to 2–3x our usage? ☐ Yes ☐ No

Notes: _____

Score (1–5): _____

Worksheet 3 : Data governance & IP ownership

Clear on where data is stored and who accesses it? ☐ Yes ☐ No

Written opt-out from training on our data? ☐ Yes ☐ No

Custom IP (code, models, prompts) assigned to us? ☐ Yes ☐ No

Clean data export & deletion on exit? ☐ Yes ☐ No

Notes: _____

Score (1-5): _____

Worksheet 4 : Security & compliance

Will conduct regular security audits and penetration testing? ☐ Yes ☐ No

Will monitor and report AI-driven decisions with full audit trails? ☐ Yes ☐ No

Will test for bias and fairness before and after deployment? ☐ Yes ☐ No

Will notify us of security incidents within a defined timeframe? ☐ Yes ☐ No

Will stay current with evolving AI regulations on our behalf? ☐ Yes ☐ No

Notes: _____

Score (1-5): _____

Worksheet 5 : Model reliability & drift control

Monitors for drift, with a retrain trigger? ☐ Yes ☐ No

Will share recent red-team / adversarial testing findings? ☐ Yes ☐ No

Human-in-the-loop on high-stakes actions? ☐ Yes ☐ No

Tool/integration controls + kill-switch + rollback? ☐ Yes ☐ No

Notes: _____

Score (1–5): _____

Worksheet 6 : Delivery, integration & implementation realism

Named, credible delivery team (not just sales)? ☐ Yes ☐ No

Has integrated with our key systems before? ☐ Yes ☐ No

Will run a POC with our real data? ☐ Yes ☐ No

Timeline survives scrutiny (no "live in weeks" for a complex build)? ☐ Yes ☐ No

Knowledge transfer / no permanent dependency? ☐ Yes ☐ No

Vendor's quoted timeline: _____ Our realistic estimate (×2–4): _____

Notes: _____

Score (1–5): _____

Worksheet 7 : Commercial & contractual

Total cost of ownership understood (build + run + retrain)? ☐ Yes ☐ No

Usage caps / alerts in place (if usage-based)? ☐ Yes ☐ N/A

Liability for AI errors defined in writing? ☐ Yes ☐ No

SLAs, exit terms, and data portability clear? ☐ Yes ☐ No

Estimated total cost of ownership (Year 1): _____

Notes: _____

Score (1-5): _____

Worksheet 8 : References & track record

Provided 3 comparable production references? ☐ Yes ☐ No

Reference call log

Reference	Real vs. quoted timeline	Support responsiveness	Surprise costs?	Would choose again?
1.				
2.				
3.				

Notes: _____

Score (1-5): _____

Red-flag check

Tick any observed. **Any tick warrants a hard second look; two or more is a likely pass.**

- ☐ Can't explain how the AI works
- ☐ Won't assign custom IP to us
- ☐ No red-team / adversarial testing evidence
- ☐ Promises full automation, no oversight, or "guaranteed ROI"
- ☐ Promises a complex build live in a few weeks
- ☐ Dismisses security / compliance / bias concerns
- ☐ Evasive on pricing or total cost
- ☐ No production references in our industry/size
- ☐ Silent on liability
- ☐ Heavy lock-in, no trial, no clean exit

Proof-of-concept plan

Before signing, define the pilot:

- Success criteria (specific, measurable): _____
- Test data set and scope: _____
- Timeline (typically 2–4 weeks): _____
- Rollback plan: _____
- Pilot cost, if any: _____

Final scorecard

#	Dimension	Score (1–5)	× Weight	Weighted
1	Strategic & domain fit		0.20	
2	Technical capability & architecture		0.15	
3	Data governance & IP ownership		0.15	
4	Security & compliance		0.10	
5	Model reliability & drift control		0.15	
6	Delivery, integration & realism		0.15	
7	Commercial & contractual		0.05	
8	References & track record		0.05	
	Weighted total		1.00	

Decision thresholds: Above 4.0 = strong candidate · 3.0–4.0 = workable if gaps are minor · Below 3.0, or any 1 in dimensions 1, 2, 3, 5, or 6 = pass.

Decision & rationale

Recommended action: ☐ Advance to POC ☐ Hold ☐ Pass

Rationale (you'll want this on file): _____

A final word

The few weeks you spend evaluating a vendor properly are cheap against the six to twelve months you'd lose recovering from the wrong one. If your AI is meant to be a source of advantage rather than a line item, the question isn't which tool to buy — it's who you trust to build it with you.

Ready to turn AI into real business results?

Tech.us builds custom and agentic AI that delivers measurable ROI.

Book a Free Expert Consultation



tech.usTM
Build Better

info@tech.us
1-866-277-3383
<https://www.tech.us/>

Tech.us empowers businesses by turning AI opportunities into measurable outcomes through innovative AI solutions catered to them. Trusted by startups and enterprises alike, we build scalable AI systems that solve real business challenges.

1,500+ Projects Delivered
25+ Years in Business
30+ Industries Served
100% Commitment to Success