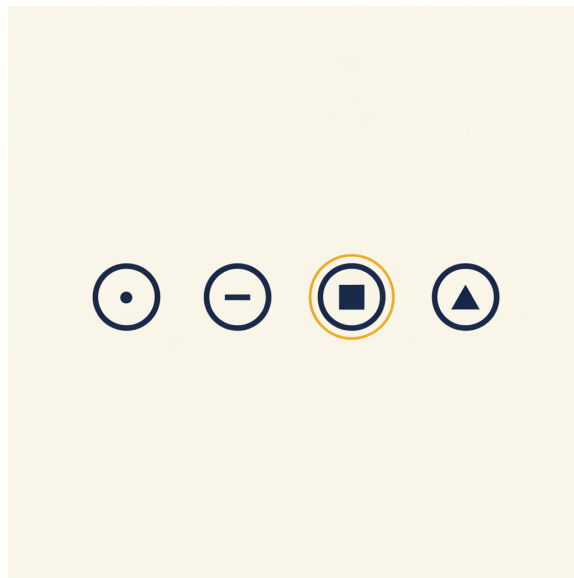




TIER 3 - EMPLOYABLE \* V1.0 -- MAY 2026

# EVALUATING AI VENDORS AND CONTRACT

Different from Tier 2's vendor research. Specifically about AI vendors -- LLM providers, RAG tools, AI consultancies, AI app builders. The questions that surface real fit vs the questions vendors love to answer.



**BY**

Alex Jahn / Agent Logic

v1.0 -- May 2026

The AI-fluent professional now responsible for picking AI vendors, AI tools, or AI contractors for their team -- and tired of marketing-fluff demos that all promise the same things

15-20 minutes

Free. Forever.

**EDITION**

**AUDIENCE**

**READ TIME**

**COST**

*Prepared by Agent Logic / alexanderjahn79@icloud.com / theaiguypi.com*

## CONTENTS

# What's in here

---

<b>1</b>	<b>The AI-vendor evaluation problem</b>	<b>3</b>
	<i>Sit through 5 AI-vendor demos in a week and you'll notice the pattern. They all show their tool's best-case scenarios. They all hit the...</i>	
<b>2</b>	<b>The 4 categories of AI vendor</b>	<b>4</b>
	<i>Different categories need different evaluation frames.</i>	
<b>3</b>	<b>The 15 questions</b>	<b>4</b>
	<i>For LLM API providers (Category A):</i>	
<b>4</b>	<b>The demo red flags</b>	<b>5</b>
	<i>? Refuses to share pricing. "We have flexible pricing tiers based on your needs." Translation: their pricing is high enough they want...</i>	
<b>5</b>	<b>The proof-of-concept structure</b>	<b>6</b>
	<i>For any AI vendor with substantial spend, demand a proof-of-concept (POC) using YOUR data on YOUR use case. Most vendors offer this;...</i>	
<b>6</b>	<b>The honest limit</b>	<b>7</b>
	<i>Three patterns to avoid in AI vendor evaluation:</i>	
<b>7</b>	<b>When no AI vendor is the right answer</b>	<b>8</b>
	<i>The under-considered option: build it yourself, or do without.</i>	
<b>8</b>	<b>Where to go from here</b>	<b>8</b>
	<i>One module left in Tier 3:</i>	

## SECTION 1

# The AI-vendor evaluation problem

---

## Every demo looks great

Sit through 5 AI-vendor demos in a week and you'll notice the pattern. They all show their tool's best-case scenarios. They all hit the same marketing beats. They all promise to "transform your operations." After the fifth one you can't remember which company did what.

This is the AI-vendor evaluation problem. The space is crowded, the demos are slick, and the differentiation between vendors is hard to surface in a 30-minute call. Teams making vendor decisions on demo-vibes alone end up with picks that don't hold up in production.

This module is the structured alternative. The specific questions that separate real fit from marketing fluff. Different from Tier 2's general vendor research because AI vendors have specific failure modes (data handling, model lock-in, capability decay) that generic vendor evaluation doesn't catch.

## What you'll have by page 12

By the end of this primer:

- The
- The
- The
- The
- The

4 categor  
15 questi  
demo red  
proof-of-c  
honest lin

***AI vendor demos hide the failure modes. The structured evaluation surfaces them. The teams making good vendor picks aren't smarter -- they're just running better evaluation processes.***

## SECTION 2

# The 4 categories of AI vendor

---

Different categories need different evaluation frames.

### The four categories:

**A -- LLM API providers** (Anthropic, OpenAI, Google, Mistral). The model itself. Evaluation: model capability + pricing + privacy posture + rate limits + SLA.

**B -- AI infrastructure tools** (RAG providers, vector DBs, observability tools, eval platforms). The plumbing. Evaluation: integration with your stack + scalability + total cost of ownership + lock-in risk.

**C -- AI applications** (sales-AI, marketing-AI, support-AI, etc.). Specific tools that do specific jobs using AI. Evaluation: fit for your specific use case + the AI quality vs alternatives + can you switch later.

**D -- AI consultancies / contractors** (services firms, freelance AI engineers, "AI strategy" providers). People rather than software. Evaluation: real prior work + ability to deliver vs sell + cultural fit + handoff plan.

Most teams make the mistake of evaluating all four with the same questions. The differentiating questions are different per category.

## SECTION 3

# The 15 questions

---

### The questions every AI-vendor evaluation should include

### Universal questions (all categories):

- 1. What's the data-handling story?** Where does our data live? Who can access it? Training-on-data policy? Retention?
- 2. What's the SLA?** What happens when the service is down? What's the credit?
- 3. What's the breach-notification commitment?**
- 4. What's the price escalation pattern over 2-3 years?**
- 5. What does the exit look like?** Data export. Contract termination. What of yours stays with us.

**For LLM API providers (Category A):**

- 6. What's the model versioning story?** When models change, is there a migration window? Can we pin to a specific version?
- 7. What's the actual latency at p95 for our query patterns?** (Test, don't trust marketing numbers.)
- 8. Caching support + pricing?**

**For AI infrastructure tools (Category B):**

- 9. How does this integrate with our existing [stack]?** Specifics, not "we have an API."
- 10. What's the migration path if we switch in 18 months?** Real answer required.

**For AI applications (Category C):**

- 11. Walk me through your top 3 customers' actual setups.** Generic case studies don't count.
- 12. What's the failure mode when your AI is wrong?** Manual review? Customer-visible? How is it caught?

**For AI consultancies (Category D):**

- 13. Show me 3 prior client projects, with reference contacts I can actually call.**
- 14. What's the handoff plan -- when do you leave, and what do we have?**
- 15. What's the project scope you'd refuse to take on?** (Tests their willingness to say no.)

The vendors who give clean, specific answers to these are the ones worth considering. The vendors who hedge or generalize are the ones to skip.

## SECTION 4

# The demo red flags

---

## What should disqualify a vendor immediately

### Demo red flags:

- ? pricing is high enough they want you committed before quoting. Walk away or push hard.
  - ? true for one customer; if true for ALL their references, they don't actually have many customers.
  - ? is suspiciously well-formatted, they may be hiding what happens with realistic data.
  - ? Either they don't know their own product, or they're hiding the problems.
  - ? Production AI adoption that fast almost always means cutting corners.
  - ? their evaluation process for switching? Vague is bad.
  - ? won't, they don't trust their tool's performance on your data.
- Three or more red flags = walk away. Two = significant caution. One = ask specifically about that one.

Refuses t

Can't sho

Demo dat

No menti

Aggressi

"We use t

Refuses a

## SECTION 5

# The proof-of-concept structure

## How to actually test before committing

For any AI vendor with substantial spend, demand a proof-of-concept (POC) using YOUR data on YOUR use case. Most vendors offer this; some you have to push.

### The POC structure:

**Timeline:** 2-6 weeks. Anything shorter is theater; anything longer is scope creep.

**Inputs:**

- 1-2 specific use cases drawn from your real work
- A test set of YOUR actual data (sanitized if needed)
- Clear success criteria -- what does "this works" look like

**Vendor provides:**

- A working setup on your data
- A specific scoreable output
- Pricing for the production version

**You provide:**

- Honest engagement (don't ghost during the POC)
- Eval criteria you'd use in production
- Internal stakeholder time for the demo back

**At the end:**

- Either you sign for production OR you don't
- Either way, you've learned something specific about whether AI tools in this category work for your problem

POCs cost money -- usually \$5k-\$20k for substantial AI tools. The cost is worth it. A bad production deployment costs 10-50x more than a paid POC.

## SECTION 6

# The honest limit

---

Three patterns to avoid in AI vendor evaluation:

- **Picking based on the strongest sales rep.** The rep is not the product. You're buying the product.
- **Picking based on the latest case study.** One impressive case study is one impressive case study. Look for breadth -- multiple cases, multiple verticals, longevity (customers staying past 12 months).

- **Picking the most "modern" or "cutting-edge" option.** Cutting-edge AI moves fast. Bleeding-edge tools today might be unsupported tools in 18 months. There's value in boring, established, stable.

Within those limits, structured AI vendor evaluation produces dramatically better outcomes than demo-vibes-driven picking. The structured process takes longer up front; the time saved over the next 2 years of using (or not using, and switching) the tool is significant.

## SECTION 7

# When no AI vendor is the right answer

---

The under-considered option: build it yourself, or do without.

**Build yourself** when:

- Your use case is highly specific and no vendor matches well
- The build is small (1-2 weeks of work)
- You have or can hire the engineering capacity
- Lock-in risk on any vendor would be high

**Do without** when:

- You haven't proven the use case actually creates value
- The AI investment is exploratory, not load-bearing
- The team isn't ready to maintain another tool

A non-trivial fraction of AI vendor purchases turn out to be unused or canceled within 6-12 months. Most of those would have been better as "wait" or "build the v0 ourselves" decisions.

## SECTION 8

# Where to go from here

---

One module left in Tier 3:

- **The AI-fluent leader pattern** -- capstone. How to advise others on AI without becoming the bottleneck.

*Get the next module the day it drops: [theaiguywi.com/training](https://theaiguywi.com/training)*

If you want this evaluation process run on a specific AI vendor selection you're considering -- the structured questions, the demo red flag check, the POC structure -- that's the consulting offer. Three weeks of structured evaluation often catches the wrong-pick before signing.

*Reach out: [alexanderjahn79@icloud.com](mailto:alexanderjahn79@icloud.com)*

## Closing -- the lock-in line

The AI vendor space is crowded and the demos all blur. The structured evaluation -- universal questions + category-specific questions + red flag check + POC -- catches the failure modes the demos hide. Teams running this process make better picks. Teams skipping it learn the hard way.

# 15

### **Fifteen questions.**

Get specific, clean answers -- vendor's worth considering. Get hedging, generalities -- skip. The structure is the difference between "we picked based on the demo" and "we picked based on real fit."

-- Alex

**Agent Logic --**

Lac, WI. This is module 17 of 18 in Tier 3 (Employable).

© 2026 Agent Logic. Share freely.

*theaiguyw*