

TEMPLATE 03 · FOR FOUNDERS BUILDING AI PRODUCTS

How to choose a design and dev partner for an AI product

Choosing who builds your AI product is the most expensive decision of your first year — and most founders make it on portfolio screenshots and a friendly call. This scorecard turns it into a comparison you can defend: nine criteria, weighted for AI products, with the exact questions to ask and the answers that should worry you. Use it on every candidate. Use it on us too.

1 hour per partner

Scorecard for 3 partners

Updated October 2026

How to use it

1. Send your [AI MVP brief](#) before the call. How a partner reacts to it is already signal.
2. Ask the questions below on the call. Score each criterion from 1 (poor) to 5 (excellent) right after, while it's fresh.
3. Multiply each score by its weight and add up. Maximum is 100.
4. **80+** strong candidate · **60–79** workable, list the gaps and ask about them in writing · **below 60** pass, however nice the portfolio is.

The nine criteria

Shipped AI products, live x3

ASK

- Show me an AI product you designed or built that real users use today. What did the model do, and what did you design around its failures?
- What surprised you after launch?

GOOD

Talks about error states, latency, cost per request, evals, how users corrected the AI — and shows it working, not in a deck.

RED FLAG

Only concept shots, a chatbot skin on someone else's API, or "we've done lots of AI" without a single live link.

Score 1–5: A ☐ B ☐ C ☐

Speed to first working version x3

ASK

- What will I be able to click at the end of week 2?
- When do real users touch it?

GOOD

A clickable prototype with real model output in week 1–2; a usable build with your data in weeks 3–6.

RED FLAG

A 4–6 week "discovery phase" with no working software at the end; a timeline that starts with "first we'll staff the team".

Score 1–5: A ☐ B ☐ C ☐

AI technical judgement x3

ASK

- For our case — would you use prompting, RAG or fine-tuning, and why?
- Roughly what will one request cost us, and how would you keep it there?
- How will we know the AI got better or worse after a change?

GOOD

Starts with the simplest approach (prompting + tools), names when RAG is needed, estimates cost per task, and mentions an eval set unprompted.

RED FLAG

Recommends fine-tuning or a custom model by default; can't estimate cost per request; "we'll test it manually".

Score 1–5: A ☐ B ☐ C ☐

Design and engineering in one loop x2

ASK

- Who writes the frontend? How does a design become code?
- Who decides when design and engineering disagree?

GOOD

Same person or a tight pair; designs in code or ships the design system as components; one accountable owner.

RED FLAG

Design and development in different companies; "the developers will figure out the states"; a handoff document as the main artefact.

Score 1–5: A ☐ B ☐ C ☐

Scope discipline x2

ASK

- What would you cut from our brief?
- What is the riskiest assumption in it?

GOOD

Cuts things, explains why, and proposes how to test the riskiest assumption first.

RED FLAG

Says yes to everything; adds features to the brief; quotes before asking a single question.

Score 1–5: A ☐ B ☐ C ☐

Pricing, contract and ownership x2

ASK

- Fixed price or time and materials? What exactly is included?
- Where does the code live, and when do we own it?

GOOD

Clear scope with a change process; code in your repository from day one; IP transfers on payment of each milestone; accounts (cloud, model API) in your name.

RED FLAG

Code in their repo until final payment; model API keys on their account; vague "phases" with no deliverables.

Score 1–5: A ☐ B ☐ C ☐

Who actually does the work x2

ASK

- Who exactly will design and build this, and can I talk to them now?
- How often will I see working software?

GOOD

You meet the people doing the work; weekly demo of the live build; overlap with your working hours.

RED FLAG

Seniors sell, juniors deliver; weekly status reports instead of demos; "the team will be assigned after signing".

Score 1–5: A ☐ B ☐ C ☐

Proof from founders like you x2

ASK

- Can I talk to two founders you built for in the last 12 months?

GOOD

Offers references without hesitation, including one where things went wrong and how it was handled.

RED FLAG

No references; only logos of big brands from years ago; testimonials without names.

Score 1–5: A ☐ B ☐ C ☐

After launch x1

ASK

- What happens when our model provider changes a model or a bug appears in month two?
- How do you hand over?

GOOD

A warranty period, a documented handover (README, runbook, prompts in the repo), an optional retainer.

RED FLAG

No answer beyond "we can quote that later"; prompts and evals not part of the deliverables.

Score 1–5: A ☐ B ☐ C ☐

Scoring sheet

Criterion	Weight	Partner A	Partner B	Partner C
Shipped AI products, live	x3			
Speed to first working version	x3			

Criterion	Weight	Partner A	Partner B	Partner C
AI technical judgement	×3			
Design and engineering in one loop	×2			
Scope discipline	×2			
Pricing, contract and ownership	×2			
Who actually does the work	×2			
Proof from founders like you	×2			
After launch	×1			
Weighted total (max 100)				

Red flags that end the conversation

- A fixed quote before they asked about your users, data or success metric.
- They can't name a single thing that went wrong on a past AI project.
- The plan has no eval set, no error states and no cost estimate — only screens.
- Your production accounts (cloud, model API, domain, app stores) would be registered to them.
- "AI will make it 10x cheaper" with no explanation of what is still done by people.
- They push their own platform or framework that only they can maintain.
- They promise accuracy numbers ("99% accurate") before seeing your data.
- Nobody on the call has shipped the kind of interface you need (chat, agent, RAG search, copilot).

Five questions for reference calls

1. What was promised for the first two weeks, and what did you actually get?
2. How far did the final cost and timeline land from the original estimate — and why?
3. What happened the first time you disagreed?
4. Could your own team take over the code without them? Did you have to?
5. Would you hire them again for the next version? What would you change?

Want this done for you?

Stepikin Studio designs and builds AI products end to end and ships MVPs in weeks. Get an estimate: stepikin.com/estimate · Book a call: vizi.club/book/aleksey_stepikin