

AI PRODUCTION READINESS CHECKLIST

25 questions to ask before your AI feature meets real users

Most AI features don't fail because of the model. They fail on what's missing around it: no way to measure quality, no guardrails, no cost ceiling, nobody accountable when it misbehaves. Answer each question honestly. It takes about 15 minutes.

How to score: 2 = yes, in place and written down · 1 = partly, or only in someone's head · 0 = no or don't know

1 • Problem and success

Score

- | | |
|--|---------|
| 1. Can you state in one sentence what the feature does for the user, and what it must never do? | ___ / 2 |
| 2. Is there a measurable success metric (accuracy, resolution rate, time saved) with a target number agreed with the business? | ___ / 2 |
| 3. Do you know the cost of a wrong answer for this use case: annoyance, lost money, legal or safety exposure? | ___ / 2 |
| 4. Is there a baseline (the current manual process or a simple rule) the AI has to beat? | ___ / 2 |
| 5. Is one person accountable for the feature's behavior in production, not just its code? | ___ / 2 |

2 • Evaluation

Score

- | | |
|--|---------|
| 6. Do you have an evaluation set of real cases (not invented examples) with expected outputs? | ___ / 2 |
| 7. Does every prompt or model change run against that set before it ships? | ___ / 2 |
| 8. Do you track quality over time, so you'd notice if it got worse after a provider update? | ___ / 2 |
| 9. Have domain experts reviewed a sample of outputs and signed off on the quality bar? | ___ / 2 |
| 10. Can you reproduce a specific bad answer a user reported (inputs, context, model and prompt version)? | ___ / 2 |

3 • Guardrails and control

Score

- | | |
|---|---------|
| 11. Is there a deterministic check around the model output (schema validation, business rules, allowed values)? | ___ / 2 |
| 12. For high-impact actions, does a human approve before anything irreversible happens? | ___ / 2 |
| 13. Are prompt injection and data leakage between users explicitly tested? | ___ / 2 |
| 14. Is sensitive data (PII, financial, health) filtered or masked before it reaches the model or the logs? | ___ / 2 |
| 15. Is there a safe fallback when the model fails, times out or returns nonsense? | ___ / 2 |

4 • Cost and performance

Score

- | | |
|--|---------|
| 16. Do you know the cost per request and per active user, and the projected monthly bill at 10x usage? | ___ / 2 |
| 17. Is there a hard spending limit or alert per day and per customer? | ___ / 2 |
| 18. Is latency measured at p95, and is it acceptable for the user experience? | ___ / 2 |
| 19. Have you tested a smaller or cheaper model against your evaluation set? | ___ / 2 |
| 20. Are you caching or reusing results where the same question comes up repeatedly? | ___ / 2 |

5 • Operations and ownership

Score

21. Are prompts, model versions and configuration versioned and deployable like code?	___ / 2
22. Is every production call logged with enough context to debug it, within your data policies?	___ / 2
23. Is there an alert when quality, cost or error rate moves outside its normal range?	___ / 2
24. Could you switch model provider within a week if prices, terms or quality changed?	___ / 2
25. Is there a written runbook: who gets paged, how to turn the feature off, how to roll back?	___ / 2

WHAT YOUR SCORE MEANS

40–50	Ready to scale. Keep the evaluation set alive and review it every release.
25–39	Ready for a limited pilot with close monitoring. Close the gaps in sections 2 and 3 before a broad launch.
0–24	Not ready for real users. The risk is not the model; it's everything around it.

Want a second opinion? I run a fixed-price **AI Production Readiness Audit**: in 10 business days I review your architecture, prompts, evaluation and costs, and hand you a written report with a prioritized plan to get to production safely. On an investment platform I led, nothing the model suggests reaches the market without human approval, with deterministic checks around every output. It has run in production for two years, with 100+ investors and over USD 100M on the platform.

David Levy · Fractional CTO & AI Engineering Lead · 20 years leading engineering teams · david@togetherdevs.net · Book a call: meetings.hubspot.com/david3299