PACKAGED AGILE
RESULTS DELIVERED

AI Capability Proof Checklist

9 Questions That Expose Fake AI Before You Buy

Compliance checks won't save you. Capability tests will.

Bring this to your next vendor demo. Ask these questions in order.

The Smoke Test

- [ ] **Is this in production with referenceable federal clients?**

  Good answer: "Yes, [Agency X] has used it for 18 months. Here's the contact."

  Red flag: "Currently in beta with select partners"

- [ ] **Does it learn from data, or follow predetermined rules?**

  Good answer: "We use supervised learning on [dataset]. The model retrains monthly."

  Red flag: Can't describe the learning mechanism (supervised, unsupervised, reinforcement)

The Wizard of Oz Check

- [ ] **What's your actual automation rate? Give me a number.**

  Good answer: "87% straight-through processing. 13% goes to human review for edge cases."

  Red flag: "Hybrid approach" without percentages

  Nate Inc. claimed 93-97%. Reality: ~0%. Result: DOJ wire fraud charges.

- [ ] **Do offshore contractors handle data, exceptions, or transactions?**

  Good answer: "No. All processing happens on US-based infrastructure with cleared personnel."

  Red flag: Vague "global support teams"

The Learning Test

- [ ] **What datasets trained the model? Do you have legal rights to that data?**

  Good answer: "Trained on [X million records] from [source]. We have licensing agreements."

  Red flag: "Proprietary blend" with no specifics

  OMB M-24-18 requires training data transparency.

- [ ] **How do you detect when model performance degrades?**

  Good answer: "We monitor accuracy weekly. If it drops below 92%, we trigger a retrain."

  Red flag: No monitoring, no drift detection, no alerts

  "Drift" = accuracy drops as real-world data changes. Good AI vendors detect this. Rule-based systems fail silently.

☐ **Why did the model make THIS decision? Show me.**

Good answer: "Here are the top 5 features that drove this prediction, ranked by weight."

🚩 **Red flag:** "It's proprietary" or can only show rule logic, not learned patterns

If they can't explain it, they can't defend it in an audit.

The Proof Test

True story: In the IRS ECM modernization, Harrison Smith required a 72-hour coding challenge using representative IRS data. Three of six vendors withdrew. Appian passed.

☐ **Will you demo with OUR data before we sign?**

Good answer: "Yes. Send us a sanitized dataset and we'll run it live."

🚩 **Red flag:** "Our demo environment doesn't support external data"

If they say no: walk away. If they withdraw: you just saved yourself a GAO finding.

☐ **Show me validated results—metrics, not testimonials.**

Good answer: "Agency X reduced processing time by 40% and error rate by 12%. Here's the case study."

🚩 **Red flag:** Only quotes, no quantified outcomes

Scoring: Count Your Checkmarks

9/9	Strong candidate. Proceed to pilot.
7-8/9	Investigate gaps before contract.
5-6/9	High risk. Require additional validation.
<5/9	Walk away. This is likely AI washing.

$5.6B	**$400K**	**70-80%**
FEDERAL AI SPEND 2022-2024 (OMB)	SEC FINES FOR AI WASHING (2024)	AI INITIATIVE FAILURE RATE

Need help running a vendor evaluation? Contact us