Brett Baden, Ph.D.
Navigar Partners, LLC

Where AI Actually Pays Off:
Validation Cost and the Case for Choosing Your First Use Cases Carefully

Many AI adoption decisions are made on the wrong question. Leadership asks whether the technology is capable enough. The better question is whether we can cheaply check its work.

Validation cost works like a transaction cost on AI output. Where checking is cheap — a compiler catches bad code, a customer repeats a request, a document fails a schema check — AI gains are larger than costs and adoption spreads. Where checking requires expert judgment and time, validation cost eats much of the gain. That single distinction explains why task-level productivity results look so strong while organization-wide results have been modest, and it gives a practical basis for deciding where to start.

The pattern

Three conditions predict where AI delivers measurable value:

1. Complementary technology is already in place. Robotics, cloud infrastructure, modern data platforms. Firms with these adopt AI at far higher rates. Two-thirds of adopting firms use AI to augment existing work rather than to replace people.
2. Tasks are bounded and verifiable. The output can be checked against a source of truth without an expert's time.
3. Skill dispersion exists to close. Gains are largest where there is a gap to close, which is why less experienced workers benefit most.

The adoption picture is more concentrated than the headlines suggest. In the Census Bureau's most recent survey, 18% of firms used AI in at least one function — 32% weighted by employment. It is not yet a general-purpose technology diffusing everywhere. It is a concentrated capability in firms that already had the infrastructure. Economic research to date has shown modest economic impact so far.

Why validation cost matters more than capability

Small failures in sequential work compound. A ten-step process at 95% reliability per step succeeds about 60% of the time, so it fails about 40% of the time. That arithmetic, not model quality, governs where agents can be trusted today. It is why AI is reliable today on limited code generation, document extraction, exploratory analysis, and tier-one service, and unreliable on anything requiring many dependent decisions.

Two things follow for planning. First, validate before you scale. The organizations getting value are the ones that designed the validation process before they deployed the model. Second, expect the ceiling to move slowly. Reliable multi-step autonomy requires per-step reliability well above 95%, and that is a capability question, not a deployment question.

There is also a cost dimension that rarely appears in business cases. AI agents do not reliably account for the computational costs they consume. Token prices keep falling, but tokens per task have risen by orders of magnitude with reasoning models and agentic workflows, so cost per operation has been rising. In one documented evaluation, an agent exhausted its allotted credits and continued anyway. Budget guardrails belong in the design, not in the retrospective.

What this means for your business

Start where validation is cheap. This is the highest-return move available. Call-center triage, document intake, invoice matching, first-pass analysis, code generation with tests. The gains are real, they arrive quickly, and — importantly — early wins build the organizational trust that later, harder deployments depend on.

Price the check, not just the build. Most business cases count implementation cost and stop. Add in the cost of validating and correcting output at the volume you expect. If a senior person has to review every output, costs quickly outpace gains.

Treat validation as a design requirement. The time-boxed access pattern matters: if an agent's access to external systems ends when its task ends, the window for unintended action narrows sharply. Guardrails are cheaper to design in than to retrofit.

Be skeptical of claims that skip the denominator. Reported productivity gains usually measure the task, not the task plus review plus rework. Ask for the version that includes review time.

Watch the horizon honestly. Long-horizon autonomy is a 2028-and-later proposition. The organizations that plan for that and capture value now on bounded work will be ahead of those waiting for the model to be ready.

The path forward

The gap between large task-level gains and modest aggregate effects should narrow as adoption spreads into more verifiable work and as validation itself gets cheaper — through better tooling, standardized evaluation, and accumulated experience about what to trust. Three things would change this view: substantial productivity gains appearing in work that requires expert judgment to check; per-step reliability on long tasks improving past 99% without heavy guardrails; or aggregate effects staying near zero even as adoption spreads into low-validation-cost work.

None of those has happened yet. Until one does, the practical guidance is the same: pick the use cases where you can check the work cheaply, design the validation before you deploy, and count the validation costs in the business case.

Brett Baden, Ph.D. is Director of Client Solutions at Navigar Partners, where he advises clients on data strategy, measurement design, and the practical economics of AI adoption. He spent more than a decade leading analytics at Nielsen and WPP/Choreograph, building predictive models and measurement systems for major consumer brands, and has published peer-reviewed research in applied economics. He holds a Ph.D. in Applied Economics from the University of Chicago. He also works with DataAnnotation, where he designs evaluation problems that probe where frontier AI models fail.

Brett Baden, Ph.D. | Navigar Partners, LLC | bbaden@navigar.partners | linkedin.com/in/brett-baden-2413951