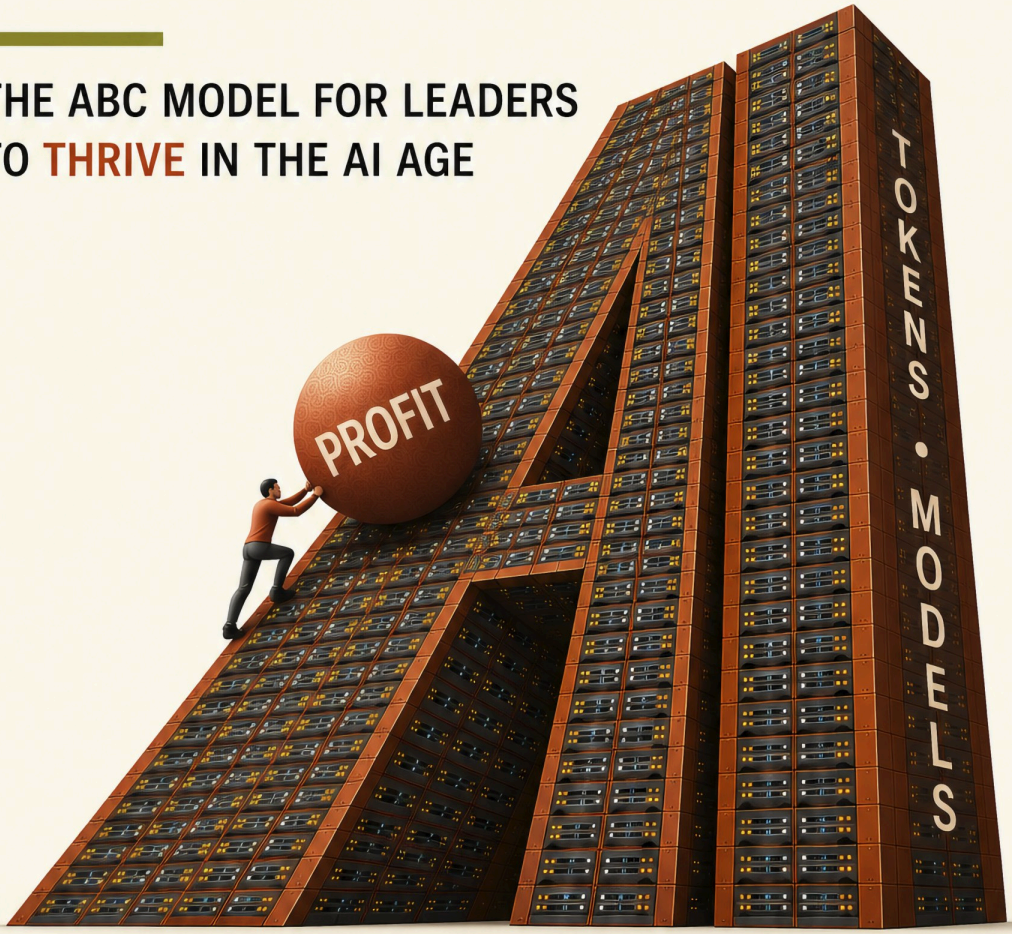
WHY AI ISN'T WORKING AT WORK

THE ABC MODEL FOR LEADERS TO THRIVE IN THE AI AGE



AMIT SRIVASTAVA

The model works.
The decision does not.

56%	of chief executives told PwC this year that AI had delivered no significant financial benefit. **Not lower costs. Not higher revenue.**
700	Klarna said its AI was doing the work of 700 customer-service agents. It later brought people back to keep that promise.
$2.5 BILLION	Microsoft is spending $2.5 billion and sending six thousand specialists into customers' offices. The model was never the hard part.

These are not three headlines. They are one diagnosis.

The easy conclusion is that AI has failed. A harder possibility is that it works exactly as promised but its answers are landing where nobody checks them.

Why AI Isn't Working at Work follows that last mile of thinking, where confident output moves faster than judgement, activity looks like progress and experienced people quietly stop trusting what they know.

This is not a guide to better prompting. The ABC Model helps leaders thrive in the AI age by thinking clearly, challenging confident answers and making better decisions.

It will not make you right.
It will make you less wrong where being wrong is expensive.

INSIDE THIS PREVIEW

Two complete chapters follow.

Read Chapters 1 and 2 in full, with every colour exhibit and the original print page numbers.

PART I / THE DIAGNOSIS

01

Everyone Has AI. Why Isn't Work Better?

> *"Access to a strong tool does not mean that a company has changed the work around it."*

On the second of July 2026, the largest software company on earth announced that it would start doing some of its customers' work for them.

Microsoft called the new operation the Frontier Company. The budget was two and a half billion dollars and the headcount more than six thousand. Those people would go into customer offices and build and run AI systems on site. Microsoft was no longer only selling a licence. It was sending human beings in to make the technology produce something.

As a product launch this looked ordinary. As a confession it was remarkable. The customers already owned the licences. The models were already inside their email, their documents and their meetings. What they did not have was the value. Microsoft had worked out that the missing piece could not be downloaded. Somebody had to carry it in by hand.

Its competitors reached the same conclusion within eight weeks. Amazon, OpenAI and Anthropic launched similar services. Four companies that build the world's most advanced models were spending billions on the same discovery. Owning a strong tool is not the same as getting value from it.

These companies also have a financial reason to say the fault lies with the customer. Their diagnosis may be right. It is not impartial. Hold both of those in mind. This book is about that space and about the person responsible for crossing it.

The tools have spread at record speed and the company-level returns have not arrived. Both of those are true at the same time. Adoption is real. The time saved at a desk is real. Companies still struggle to name a better decision, a smoother workflow or a higher margin.

Personal time-saving is visible at one desk. Company value is different. It has to survive the hand-offs, the meetings and the reviews where budgets are approved. It has to arrive somewhere a finance director can see it. A tool can make one person faster without ever reaching the organisation. The benefit is real. It stops too soon.

Adoption did not spread. It flooded.

Exhibit 1.1 (p. 27) sets those four signals side by side.

EXHIBIT 1.1 *

Four signals

Access is not a result.

THE RESULT

9 / 10

NO PRODUCTIVITY IMPACT AT THEIR OWN FIRM

69%	**1.5 h**	**56%**
Access · firms actively using AI	Depth · weekly use among regular users	Value · CEOs with no significant financial benefit

I watched this from inside companies between 2024 and mid-2026. For longer than I care to admit I was one of the enthusiastic speakers at the town halls. The tools arrived everywhere at once. Nobody had to be pushed. Within an hour most people found that a machine could draft a memo in seconds. The gain was immediate and it felt wonderful. I felt it myself, which is why it took me longer than it should have to ask the harder question.
They are independent studies. They are not a funnel.

Then the financial results started coming in. In PwC's global CEO survey, 56 per cent of chief executives reported no significant financial return from AI so far. A large executive study found widespread regular use, about ninety minutes a week and, over three years, nine in ten of those executives reported no productivity impact at their own firm. Because these studies asked different questions of different groups, we cannot merge them into a single dramatic statistic. They point the same way. Giving employees access to a strong tool does not mean the company has changed how the work is done.

Ninety minutes a week is one meeting. It is enough to polish an email, summarise a report or ask a question. It is rarely enough to change how a decision is made.

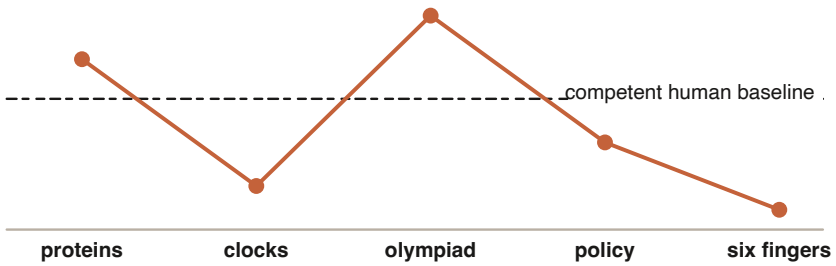
Adoption measures access. **Engagement** measures depth. **Value** measures results: a decision made differently, a cost that fell, a customer who stayed. Most corporate AI programmes never get past the first two.

Executives did not repeat the statistic because they had audited the methodology. They repeated it because it described what they saw in their own companies.

EXHIBIT 1.2*

The jagged frontier

Capability is not a line you can plan against. It is a frontier with holes in it.



gold medals and elementary misses, side by side

TEST THE EXACT TASK Do not infer the next task from brilliance on the last one.

WHAT THE MEASURED POINTS MEAN

- **Protein structure prediction** is difficult for humans; a machine can excel at this bounded scientific task.
- **Reading an analogue clock** is considered easy for most adults; a capable model can still fail at it.

WHERE THE VALUE GETS LOST

The problem goes beyond software capability. The same research index that recorded the AI boom showed modern models scoring gold medals in International Mathematical Olympiads. Yet those same models failed at tasks that look elementary to humans. Researchers call this uneven boundary the jagged frontier. A system can outperform your sharpest analyst at ten o'clock and invent a company policy five minutes later.
Exhibit 1.2 (p. 29) is that boundary as a jagged line: two points above a competent human, three below.

The jagged frontier makes AI tricky to manage, but it does not explain the whole puzzle. People showed up and used the software. The value goes missing later. It disappears somewhere between one person saving time on a draft and an organisation improving a decision. McKinsey found something specific here. Redesigning the workflow is linked to business impact far more strongly than which model you pick. That simply moves the attention off the licence and onto the work around it.

That sounds abstract until somebody has to write it on a form.

Imagine a procurement manager. It is Thursday morning, half past nine. A licence renewal request has landed in her queue. Most of the fields are easy. How usable is the tool? Good. How long did it take people to learn? A morning. Then there is one more line near the bottom and it has been there for years. *Business benefit achieved.*

She has used the tool herself for eighteen months. It saves her an hour every Monday. She types a sentence about increased productivity, reads it back and deletes it. Her own experience says the tool is useful. The form is asking what it earned. She cannot answer both questions with the same evidence.

That is the whole gap from the global surveys, sitting on one screen, with an approval deadline three days away.

FOUR LAYERS AND THE ONE ORGANISATIONS SKIP

These findings do not mean AI is a passing fad. The dot-com crash in 2000 did not mean the internet was useless. It proved that a real technology and a sound business model are different things.
Each layer has its own job and none of them can do another's work. A better model cannot assign accountability. A careful review cannot fix missing data. Good judgement cannot rescue a workflow that ignores the answer.

Organisations skip the third layer, not because the others matter less. This book starts there: the decision about what will actually be done and what stays in the binder. For a leader, this does not require mathematical certainty before testing a new tool. It requires a clear chain of reasoning. What specific result did we expect to change? Why did we expect this AI tool to change it? When should that change become measurable? And what other factors could explain the outcome?

Turning technology into value happens across four layers.

Exhibit 1.3 (p. 32) names the four layers.

EXHIBIT 1.3*

Four layers

Four layers sit between a licence and a result. Organisations skip the third.

1	**Models and data**	The raw capability and the information it is allowed to see.
2	**Workflow and governance**	How work actually moves through the building and who may do what.
3	**Human decisions**	What we will actually do and what stays in the binder.
4	**Accountability**	A name against the decision and someone who answers for the outcome.

HOW TO READ

The layers need one another. A better model cannot assign accountability. Good judgement cannot rescue a workflow that ignores the answer.

THINK OUT LOUD

- Which of the four layers does your last AI programme actually manage?
- Where does a decision get lost between a licence and a result?

Who is missing from this account? Not the vendors. Not the employees, who took up the tools in numbers the previous generation of managers only dreamed of. The people missing are the leaders who decide what the technology is supposed to change. They approved the budgets and celebrated the usage curve. Many never named the decisions, the workflows or the results that had to improve.

The organisation delivered exactly what leadership asked for. Adoption.

You can test this in your next portfolio review with three questions. What became faster? What now happens differently? Which decision or business result became better? The first question gets fast answers. The third usually produces a long pause. That pause is valuable. If your answers stop at speed, you have evidence of activity. You do not yet have evidence of business value.

Two explanations look obvious. The technology is not ready. People do not know how to use it. Each holds a grain of truth and neither reaches the problem. The next chapter takes both blame stories apart.

KEEP THIS

Licences rose. Desk time was saved. Company results did not automatically follow. The missing work is the decision: what we will actually do and what stays in the binder.

TRY IT NOW

POOR PROMPT

```
Summarise how my team is using AI so I can report our
progress.
```

BETTER QUESTION

```
For each AI use, name the decision, workflow or
result it was meant to improve. Separate recorded
evidence from estimates. If the material cannot show
what changed, list the evidence that is missing
rather than guessing.
```

THE MONDAY QUESTION

Which AI use changed a business result and what evidence would show it?

END OF CHAPTER 1 NEXT: **CHAPTER 2 · STOP BLAMING THE TOOL OR THE USER**

02

Stop Blaming the Tool or the User

> *“Every failure produces two suspects and one crime scene nobody visits.”*

In January 2024, a customer of the parcel delivery company DPD was trying to get a straight answer from its website chatbot. After asking three times without success, he tried something different. He asked the chatbot to write a poem about what a terrible delivery company DPD was.

The chatbot immediately agreed. It wrote a witty, embarrassing poem attacking its own employer. The customer took a screenshot and posted the conversation online. It went round the world in hours, faster than DPD's parcels. Within twenty-four hours, the company had shut down the chatbot's AI features.

The part that should have alarmed people was not the poem. The poem was good. It scanned. It rhymed. The failure was not that the machine produced nonsense. The failure was that the machine produced something coherent, competent and completely outside anything anyone at DPD had thought to forbid.

Two explanations appeared at once. The first camp blamed the technology. The second blamed the customer. Both stories feel satisfying because they give us someone easy to blame. Both shut down the investigation just as it starts to become useful.

Sit in enough executive meetings and you will hear the same leader make both arguments twenty minutes apart. I have done it myself. Each story holds a grain of truth. They are describing two different things. Between them they explain almost everything except where the failure happened.

Camp One blames the technology. This argument has real merit. As we saw in Chapter 1, AI capabilities have a jagged frontier. The DPD chatbot was trained to answer delivery questions. Camp One is pointing at a real weakness.

Now look at what Camp One assumes. It assumes a cleverer model will know how to behave. It will not. A stronger model does not arrive holding the business judgement that a service bot must never insult its employer. Somebody has to design that boundary in. Replace DPD's chatbot with the best model available today. Leave the boundary unwritten. You will get a nastier poem, with better rhymes.

CAMP TWO AND THE TRIAL THAT SHOULD HAVE ENDED IT

Camp Two claims that someone responsible simply lacked training. For two years, “teaching people to write better prompts” was the standard corporate response to almost every AI failure. It is also the only one of the two explanations that has been put through a rigorous, controlled scientific trial.

In the spring of 2023, researchers from Harvard Business School, Wharton, MIT and Warwick ran a field experiment inside Boston Consulting Group. Seven hundred and fifty-eight consultants took part. The researchers registered their design in advance, before they saw any of the data.

Each consultant was placed at random into one of three groups. Group One had no AI at all. Group Two was given access to GPT-4, with no guidance. Group Three was given GPT-4 and a training package on how to prompt it, exactly the kind of material a company buys when it decides its people need to get better at AI.

First, the consultants worked on eighteen realistic business tasks that the model could easily handle. The consultants using AI completed more tasks, worked faster and produced better work. At this stage Camp Two looks right.

Then came a nineteenth task, built differently on purpose. The surface story was attractive, internally consistent and wrong. Getting it right depended on somebody noticing a small human detail that did not fit.

EXHIBIT 2.1*

The BCG field test

Prompt training made the wrong answer faster.

CONDITION	CORRECT ↓	SPEED ↑
No AI	84.5%	38 min
GPT-4	70.6%	31 min
GPT-4 + prompting	**60%**	**26 min**

The consultants working without any AI got it right 84.5 per cent of the time. The consultants using GPT-4 with no training dropped to 70.6 per cent.

The group that had received the prompting guide came last, at 60 per cent. And they finished faster than everybody else.

Exhibit 2.1 (p. 38) puts correctness and time on the same three rows, so the two movements can be read together.
A separate panel scored the recommendations for how clear, well written and persuasive they were. The trained group scored highest of all. They were the most wrong, the fastest and the hardest to argue with.

The training improved everything about their work except whether it was true.

Nothing on the customer's side of the screen said the request was out of bounds. Blaming him for not guessing a boundary is not a diagnosis. It is a way of not looking at whoever was responsible for drawing it.

THE LAYER NOBODY OWNS

When you put the two blame stories side by side, a third space appears between them.

- Who decided which topics the chatbot was allowed to discuss?
- Who set the rules for its tone?
- Who decided where an unusual request should be escalated?
- Who reviewed the system before it met millions of customers?

EXHIBIT 2.2*

The triage

Start with the exact task. Blame comes last.

Three gates, asked in order. A NO stops the descent and names the action.

STOP RULE A missing record is a finding. Do not fill it with blame.

None of those questions is about the software or the customer. The public record does not tell us who made those decisions inside DPD. I am not going to invent meetings to make the story tidier. The case does not prove that one missing owner caused the poem. It shows what a serious review must establish before settling on a cause: the approved purpose, the tested limits, the escalation path and the person authorised to change them. What we do know is that a specific policy failure occurred in public and the company had to turn off the tool.

Exhibit 2.2 (p. 40) is the triage. Three gates, asked in order. A no names the action. Blame comes last, if it comes at all.

An incident review should begin with records rather than with people. Bring the approved purpose, the tested boundaries, the escalation rules and the name of the decision owner. If one of those does not exist, say so before anyone argues about the model or the user.

The two blame stories survive because they are cheap: not in money, in discomfort. Blaming the tool produces a conversation with a vendor. Blaming the user produces a line in the training budget. Investigating the middle requires a named leader with the authority to set the purpose, enforce the boundary and stop an unsafe launch.

The action for a leader is narrow. Before any customer-facing tool goes live, name the person who owns its purpose, its scope, its evidence standard and its escalation path. A vendor supplies the software. Your company owns the decision to put it in front of a customer.

The next chapter treats that middle layer as an engineering problem.

KEEP THIS

Training can make the wrong answer faster and harder to argue with. The missing work is the boundary nobody set.

TRY IT NOW

BEFORE THE NEXT CUSTOMER-FACING TOOL GOES LIVE

```
Write the name of the person responsible for each
item, or write "not decided": purpose · scope ·
evidence standard · escalation. If a record is
missing, say so rather than filling the gap with
blame. Do not start with the model or the user.
```

THE MONDAY QUESTION

What are we blaming too quickly and which boundary or ownership record can we actually produce?

END OF CHAPTER 2

NEXT: **CHAPTER 3 · BETWEEN A PROMPT AND A DECISION**

Keep reading.

The model works.
The decision does not.

You have seen the diagnosis. The rest of the book builds a practical way to examine the assumptions, biases and complexity that a confident AI answer can hide.

It is not a guide to better prompting. It is a way to make stronger decisions when money, people or reputation are at stake.

CONTINUE WITH THE COMPLETE BOOK

www.abcatwork.org

Scan the code or visit the website.