insightology.ai

Build, Buy or Borrow?

Choosing your agency's AI strategy

A working guide to deciding between building your own AI, creating skills on a generic AI assistant, and using a specialist platform

A WHITE PAPER BY INSIGHTOLOGY

JULY 2026

This is a strategy decision, not an IT purchase

01

A quick word on the title. Build and buy explain themselves: build means coding your own specialist AI platform, and buy means renting one from a vendor. "Borrow" is the third path: borrowing the power of a generic AI assistant (an engine you don't own or control) and adding your own layer of research skills, prompts and agents on top. The distinction from buying matters: you buy into a specialist platform, methodology included; you borrow a general engine and supply the research knowledge yourself. Borrowing is sometimes exactly the right answer, and this paper will tell you when.

AI is resetting the economics of research. Production work (proposals, survey design, fieldwork materials, first-pass analysis) can be done far faster, agencies that reprice can earn higher margins, and demand is growing at the strategic end of the market. **This paper is about the decision that follows: what kind of AI capability your agency should run on.** Framed properly, that is a decision about machinery for selling warranted thinking (judgement a client can trust, from a process a client can check), because whichever path you pick is the thing that will, or won't, let you prove your work.

Adoption is no longer the question. The Marketing Research Institute International's April 2026 survey of 498 insights professionals found that around nine in ten now use AI in some capacity, but only 13% say it is fully embedded in how they work. What agencies do next is a strategic decision, not a procurement one, for three reasons:

It sets your cost base

> The AI path you pick determines whether AI is a fixed overhead, a per-seat subscription, or a cost that moves with the work you win. Those are three different businesses.

It sets your ceiling

> A chat assistant caps out at helping individuals go faster. A specialist platform changes what the whole agency can take on. A build, done well, becomes a product in its own right. You are choosing how much upside is available, not just how much you'll spend.

It's hard to reverse

> Eighteen months into a build, or two years into retraining ninety people on a toolkit, changing course is expensive and demoralising. The cheap time to think hard is now.

This edition updates the July 2026 paper. It adds the two things agency leaders most often ask us for: a clear picture of what a skills layer actually is and how it works, and the evidence on how generic AI performs on professional-grade work, summarised in Section 4 and set out study by study in the appendix.

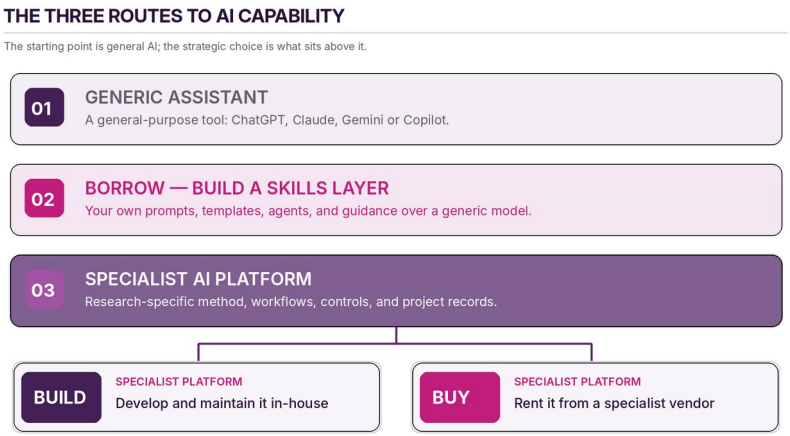
The usual disclosure: Insightology sells a specialist AI platform to research agencies, so we're not neutral bystanders here. We've written this to be useful even so, including being clear about when the right answer isn't us. Hold us to that.

The question is never "which option is cheapest?" It is "which option wins us more than it truly costs us to own?" Keep that in your head for the pages that follow and the decision mostly makes itself.

The three options, properly described

02

We have met hundreds of market researchers, and each is using a different mix of AI tools. For our industry they fall into three kinds of technology, and the third comes in two commercial forms. They are as different as chalk and cheese, which does not make the differences easy to see. Here is the map, and how the title of this paper lines up with it.



The starting point: a generic assistant

ChatGPT, Claude, Gemini or Copilot: a chat window on a foundation model that has read most of the public internet. A brilliant generalist, a mile wide and an inch deep; for tactical, low-stakes work it may be all you need. Its limit is the definition: it produces the most likely-sounding answer, its output varies run to run, and nothing tells you when it is wrong.

Borrow: a skills layer on a generic assistant

No software gets built. You write your agency's ways of working into instructions the assistant follows: prompts, templates, reference documents, agents, training, a governance policy and a senior owner of quality. Most agencies try this path first, and it is the most misunderstood, so Section 3 opens it up.

Build: your own specialist platform

Designing and coding a specialist AI system yourself, with your methods and IP engineered in, on foundation models rented from the big AI labs. You own the product, and it might become sellable. It is also a change of business model: you would be running a software company alongside a research agency.

Buy: rent a specialist platform

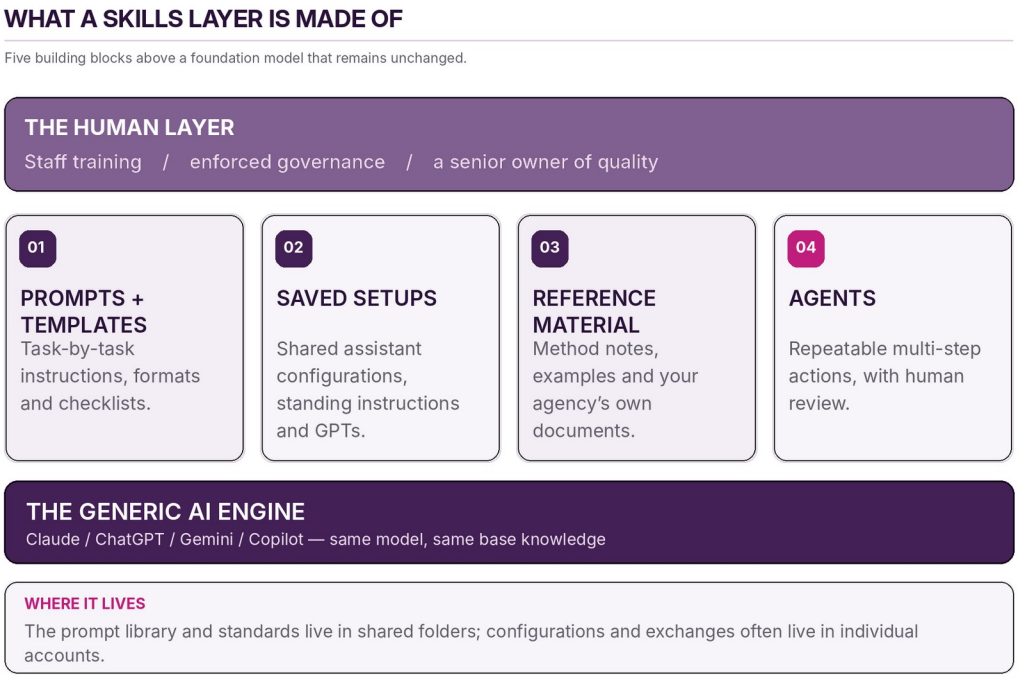
The same kind of system, built for research by a vendor: methodology, workflows, project record, audit trail and certifications arrive as a finished product. Your side is smaller but real: onboarding, your house methods, connecting your tools, and changing how projects run.

A useful shorthand. A generic assistant is buying a tool. A skills layer is buying a productivity boost. A self-build is becoming an AI engineering company. A rented platform is renting that capability without becoming one. The right answer depends on which your agency needs, and which it is equipped to run.

What a skills layer actually is

03

Because no software gets built, it can be hard to picture what you would actually own at the end of a skills-layer programme. It is worth being concrete, because its shape explains both what it does well and where it runs out. There are five parts, and they all sit on top of a generic engine that does not change.



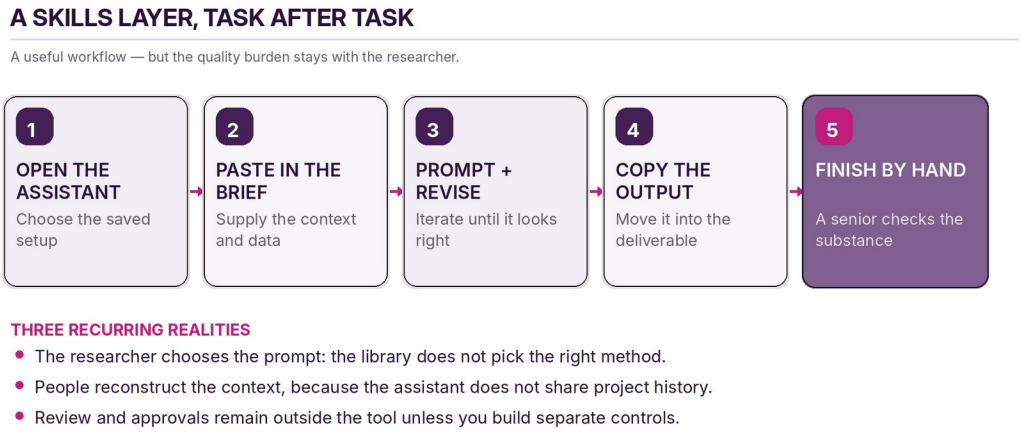
Notice who has to write most of it. Agencies often hand this work to the most AI-literate people in the building, who are usually the most junior. It is the wrong way round. A skills layer is only as good as the expertise written into it, so the job belongs to the people with the most experience to codify. Hand it to juniors and you codify the wrong level of know-how, and the layer is too weak to trust.

There is no system to log into. The interface is the same chat window your people already use, and the outputs are the same documents a generic assistant produces. Larger set-ups manage the library centrally, with shared prompts, permissions and logs, and that is still a skills layer: the method remains outside the engine. Much of a mature layer ends up in individual accounts, where the agency cannot see it.

A word on agents

Agents are the part people ask about most, and the part most often oversold. An ordinary prompt gives the model one instruction and gets one answer back. An agent gives it a goal and permission to take several steps on its own: read the brief, pull the template, draft each section, check it against a checklist, hand back a document. For repeatable work with a stable shape (reformatting, extracting, summarising to a fixed structure) they are genuinely useful, and the most valuable part of a good skills layer. Three things are worth understanding before you build a lot of them.

- **An agent inherits the engine it runs on.** It adds reach, not knowledge. If the model improvises a methodology, an agent improvises it faster, in more places, with less chance of anyone noticing.
- **They are brittle.** Each step assumes the last returned something in the expected shape. A small drift early on compounds, and the failure often appears several steps later, somewhere that looks unrelated.
- **They are hard to see inside.** When a five-step agent produces something wrong, finding the step that failed is real work, and this is usually when an agency discovers nobody wrote down how the agent was built.



Four everyday tasks show the pattern.

Writing a proposal

Open the proposal setup, paste in the brief, and a full draft arrives in minutes. The structure and language are good. The methodology thins out: plausible, generic and not obviously wrong, which is the difficulty. It cannot know which of your six similar past studies is the right precedent, and it will fill in timings and costs if you let it.

Designing a survey

A sensible draft: sections in a reasonable order, routing notes, wording that reads well. It misses what actually breaks studies: whether the scale will discriminate, whether the order primes the next answer, whether the instrument answers the analysis plan. None of that surfaces until fieldwork or analysis, when it is expensive.

Analysing multi-modal data

Uploading confidential fieldwork to a generic tool raises security and governance concerns. Even where permitted, a general AI model can misread figures, sample bases, respondent details, transcripts and quotations, then confidently summarise the wrong thing. Its general training is no substitute for analysing your evidence with a validated research method.

Across batches of transcripts, open-ends and prior waves, the coding frame can drift. Ask twice and themes may change, without a reliable trail back to the respondents. Every finding still needs checking against its source.

Building the debrief

A well-organised narrative with confident headlines. What it cannot supply is the part that matters: which finding bears on the decision the client is about to take, what to lead with when two stakeholders disagree, and which recommendation you will defend. It writes the deck. It does not decide what the deck should say.

Read those four together and a pattern appears. **A skills layer is reliably good at the shape of the work and unreliable about the substance of it, and the substance is the part a client is paying for.**

A brilliant graduate with your manual

Leaders hear "AI with research skills" and picture a research expert. What they have is closer to a brilliant generalist with a research reading list. Imagine hiring the brightest graduate you have ever met and handing them your agency's operating manual. They study it for a fortnight and could recite it back to you. Then a brief arrives for a needs-based segmentation of adults with poorly controlled type 2 diabetes. You would not expect their first-pass questionnaire to come back more than half right, and you would not send it to the client without a lot of senior crafting. Not because they are not clever, but because they have never designed a survey like that, watched it run in field, and watched it come apart at analysis because the needs statements did not separate one segment from another. That sequence is where the expertise lives, and none of it is in the manual.

> **The manual holds your method. It cannot hold the experience of having run it.**

A generic assistant loaded with your prompts is that graduate. We have direct experience of this, because the first AI system we built was an advanced skills layer, and we tested it over and over. Even with a detailed instruction manual and a library of strong examples behind it, a substantial share of its outputs contained material that was invented: not drawn from anything we had given it, and not following any research logic we would recognise as sound. The uncomfortable part was not how often it happened. It was that the invented material looked exactly like the rest. It walked like a duck and it quacked like a duck. It was not a duck.

Set out plainly, here is what a skills layer on a generic assistant is, and what it is not.

What a skills layer is	What a skills layer genuinely is not
A brilliant generic assistant. With good prompts and skills, a fast producer of first drafts: guides, summaries, desk research.	**A research expert.** It improvises from general training, so quality varies run to run and person to person, with no QA layer underneath.
A procedure-follower. Ask it to apply a method and it will have a competent go.	**Able to warrant the result.** A subtly wrong output looks identical to a right one, and the only workings you can show a client are a chat log.
A personal memory. Modern assistants remember an individual's context and preferences.	**An agency system of record.** No shared project history, approvals or audit trail; project 200 learns nothing from the 199 before it.
Advisory. It does whatever the person typing asks, in whatever order they ask it.	**An enforced workflow.** Steps can be skipped or reordered, so rigour depends entirely on who is at the keyboard.
Acceptable for internal, non-confidential work. Enterprise tiers add real data protections.	**Something most security assessments will accept for client or respondent data**, especially as teams actually use it, on personal accounts.
Improving at astonishing speed, for everyone at once, including every competitor you have.	**Yours.** Nothing about your usage accumulates as a capability the agency owns.
Less expensive than coding your own specialist system.	**As cost-effective as you might think**, and sometimes dearer than renting (Section 6).

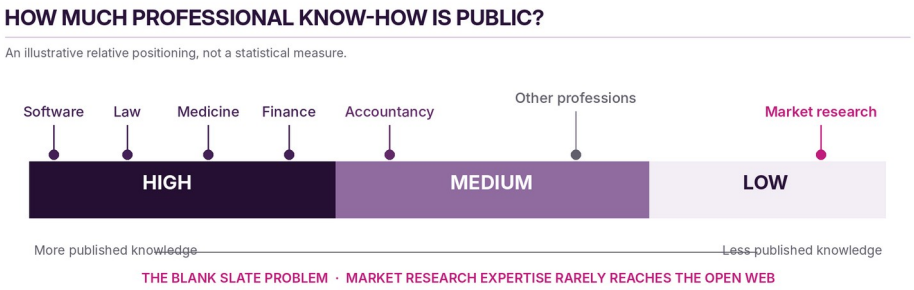
Understanding the limits of generic AI and skills layers on professional-grade tasks

04

None of this makes a skills layer a bad idea. For simpler, repetitive tactical work, where a senior has time to review and improve the output, it can be the right answer. The question is what happens on professional-grade work: the strategic projects where an output has to be right and you have to be able to show that it is. Three things decide that: what the model was trained on, the costs that never reach an invoice, and what independent studies have measured.

Market research is a hard industry for generic AI

Generic models train mostly on public data. In law, accountancy, medicine and software engineering, vast bodies of professional know-how are published: case law, statutes, filings, clinical literature, millions of lines of open-source code. A generic model arrives in those professions already steeped in the raw material. There is no equivalent public library for market research.



Three structural reasons keep our knowledge off the open web, and none of them will change. **Confidentiality:** proprietary methods, client deliverables and panel data sit behind NDAs by contract. **Tacit craft:** a large share of high-level research skill has never been written down by anyone; it is pattern recognition built over decades and passed on by apprenticeship. **Unpublished fieldwork:** transcripts, data files, open-ends and discarded hypotheses rarely leave the project folder. Together they create what we call the blank slate problem. On the subjects where our profession's real expertise lives, a generic model arrives knowing relatively little, and answers in exactly the same confident tone whether it is standing on a mountain of training data or on nothing at all.

Ask it what conjoint analysis is and the answer is excellent, because the textbooks are online. Ask it to specify one for a real study and nothing published settles how many attributes before respondents satisfice, whether price should be conditional on brand, or what to leave out because the client cannot make it. It improvises, in the same even tone. And prompting does not close the gap: **prompts direct; they do not educate.** A model will follow a fifteen-step method, but its grasp of what each step is for, and what quietly goes wrong at step nine, comes from its training. You cannot prompt twenty-five years of fieldwork into a machine.

Three costs that never reach an invoice

01

Writing the manual

A skills layer is your methodology written down: not bullet points but how a screener is structured and why, what a guide must achieve at each stage, what good looks like at every checkpoint. Most agencies have never written it down. That is director time, a great deal of it, from the people whose hours you hoped to free. And it has to be right: a mistaken instruction produces every output flawed the same way.

02

The polishing tax

Senior hours spent lifting almost-right output to a professional standard. A first draft that is 60% right is not a 60% saving; you have added an editing job on top of a thinking job. Several agency leaders have told us it was quicker to write a screener from a blank page than to fix what the chatbot gave them. If a draft is not close to client-ready, you do not save time. You move it to the level above.

03

The rebuild cycle

General models are replaced every few months, and a serious prompt and agent library has to be re-tested and updated each time. The version you tested in March is not the version running in September. Most agencies discover this in year two, and each rebuild sends them back to the first cost, because the manual has to be checked against a model that behaves differently.

And one quieter limit, which has nothing to do with quality. **A skill is knowledge, and knowledge diffuses.** The prompt library that gives you an edge in March is on a competitor's laptop by June, because there is nothing to it except words anyone can copy. **A specialist platform is infrastructure, and infrastructure accumulates.** Keep that pair in mind through the economics that follow.

What the research shows

The polishing tax is not a theory of ours. It has been measured in software engineering, office work, law, tax and finance. Every profession studied is one the models know well, so these results were gathered on the most favourable ground a skills layer will ever have. Here are the six findings first; the individual studies are presented in the appendix, including the findings that are awkward for our argument.

01

Fluency is not reliability. Outputs from generic tools sound plausible, but hallucination, poor quality and quality constraints remain common, so every output still needs detailed checking.

02

The first draft decides the saving. The polishing tax is the normal shape of AI on an unchanged process, and it is usually paid by the most senior person in the chain.

03

A skills layer has been measured on finished work, and it fails much of it. Independent tests of leading general models show that prompts and instructions improve drafts, but they do not reliably enforce method or remove the need for expert review.

04

Any advantage that lives in the model lasts about eighteen months. The durable edge now sits higher up: in the workflow, the sources, the record and the trail.

05

Specialists hold their edge on grounded, checkable production. Where they have been independently evaluated, specialist systems using workflow, sources and audit trails beat the professionals they were set against on defined tasks.

06

Every measurement was taken on the models' best ground. For a skills layer in market research, read these results as a ceiling, not a floor: our data is more private, less standardised and easier for a generic tool to misread.

What makes a platform specialist

05

Specialist AI means a system designed around the needs of one profession; you will also see it called domain-specific or vertical AI. It is not a generic assistant with better prompts. It is a different kind of thing, with five distinguishing parts. If a tool in front of you is missing several of them, whatever the label says, it is a skills layer.

1. **Knowledge beyond the public web.** Validated methodologies and expert-designed workflows, designed by seasoned researchers and coded by AI engineers, plus a secure store of your agency's own best work, uploaded as the documents you already have.
   The test: What does it know that the open internet does not, and can our own know-how live inside it?

2. **Method in the machine.** Workflows run brief to debrief, stages feed each other, checks are enforced, and steps cannot silently be skipped or reordered.
   The test: Can a step be skipped or diluted by whoever is at the keyboard?

3. **A system of record.** Every stage of every project is remembered, so project 200 benefits from the 199 before it, and that benefit belongs to your agency alone.
   The test: What does it know about our agency this month that it did not know last month?

4. **An audit trail.** What was run, on what data, amended and approved by whom. When a client challenges a finding there are workings to show, and a senior can audit how a junior built an output before it reaches the client.
   The test: If a client challenged a finding tomorrow, what could we put in front of them?

5. **Security and compliance designed in.** Access controls, data residency, retention and deletion, a contractual and architectural guarantee that client and respondent data trains nothing, and SOC 2 and ISO 27001 carried by the system.
   The test: Can it show where respondent data goes, and will a client's procurement team accept its certifications?

Put the five together and what you get is not a smarter assistant. **It is the profession's rulebook, delivered as software.**

Comparing our own platform against the skills layer we built first, the platform produces more accurate outputs from brief to debrief, especially on complex tasks, and they need far less editing before they are client-ready. Other platforms and other skills layers will differ; that is our experience, not a universal rule. But it matches the independent studies in Section 4.

One common misreading. People sometimes assume a specialist platform competes with the frontier models, so that if OpenAI or Anthropic get good enough it becomes redundant. It is the wrong picture. The models supply raw capability; the specialist layer supplies the validated method, the enforced process, the record and the trail. They sit on top of each other, which is why a good platform supports several models and swaps between them as the frontier moves, and why model upgrades generally make the layer above more valuable, not less.

Side by side

WHERE THE EXPERTISE SITS

SKILLS LAYER

In written instructions, applied by whoever is typing

SPECIALIST AI PLATFORM

In the engineering of the software

RESEARCH KNOWLEDGE

SKILLS LAYER

Whatever reached the public web, plus your written manual

SPECIALIST AI PLATFORM

Validated methods and research logic built in, plus your own private canon, connected to the workflow

KNOWING WHEN TO STOP

SKILLS LAYER

Improvises fluently where evidence is missing, without flagging that a decision is needed

SPECIALIST AI PLATFORM

Stops where the method says there is a decision, and says what is missing

BEING RIGHT, AND PROVING IT

SKILLS LAYER

A subtly wrong output looks identical to a right one; the only workings are a chat log

SPECIALIST AI PLATFORM

Validated methods and an audit trail from brief to debrief

WHAT ACCUMULATES

SKILLS LAYER

Very little: each session starts again from close to zero

SPECIALIST AI PLATFORM

Project 200 benefits from the 199 before it, for your agency alone

SECURITY AND COMPLIANCE

SKILLS LAYER

Yours to carry; consumer versions fail vendor assessments, enterprise tiers cover part

SPECIALIST AI PLATFORM

Built in, with certification carried by the vendor

ANNUAL COST, THIRTY PEOPLE*

SKILLS LAYER

About £105k a year once running, £130k in year one

SPECIALIST AI PLATFORM

About £60k all in, scaling with projects won

GENUINELY GOOD FOR

SKILLS LAYER

Tactical production work and first drafts

SPECIALIST AI PLATFORM

Strategic work you have to be able to warrant

See Section 7 for the cost lines and assumptions

The profession that ran the experiment

06

Commercial law has already faced the choice market research agencies are making now: rely on generic AI with prompts and agents, or invest in systems designed around professional work. It is an unusually useful comparison. Both professions sell judgement under uncertainty; both must defend their work to clients; and in both, a subtly wrong result can look perfectly sound until real money is at stake. Law also began with an advantage we do not have: the models had already read a vast public library of cases, statutes and commentary.

What law firms chose — and why

Many law firms started with general assistants. Increasingly, leading firms adopted specialist legal platforms, including Harvey, which reported serving more than 100,000 lawyers across 1,300 organisations in 2026. The attraction was not simply better drafting. It was professional assurance: checkable answers grounded in appropriate sources; contractual protection for privileged client material; and controlled workflows with review, approval and a record of what was done. In other words, the profession wanted more than a convincing document. It wanted work it could stand behind.

What the evidence tells us

A 2025 randomised study at the universities of Minnesota and Michigan compared a specialist legal tool, a general reasoning model and no AI on realistic legal tasks. Both AI systems increased productivity substantially. The more revealing difference was in invented citations: eleven with the general model, three with the specialist tool. Later agent benchmarks tested whether general models equipped with tools could complete finished legal work. On the stringent all-criteria measure, even the strongest passed roughly one task in four. Meanwhile, independent evaluations of Thomson Reuters' CoCounsel reported scores of 73–90% across four legal task types, each more than ten points above a practising-lawyer baseline. Different studies measure different things, so none proves that every specialist wins every task. Together, they show why fluent output and professionally defensible work must be judged separately.

The test that mattered most

In 2026 Anthropic entered legal work directly, first with a legal plugin and then with Claude for Legal, offering connectors and practice-specific tools. If specialist platforms were only prompt libraries wrapped around a foundation model, this could have displaced them. Instead, specialist providers including Harvey, Thomson Reuters, LexisNexis and Legora participated in the new ecosystem. The shape of the market became clearer: frontier models provide raw capability, while specialist systems provide the sources, rules, workflow, security and accountability needed to put that capability to work. Commercial adoption alone is not proof of performance, but it shows which kind of system major firms were willing to rely on.

The lesson for market research

The analogy has limits. Major law firms have budgets and regulatory obligations few agencies share, and some legal tasks are more repeatable than strategic research. Yet the most important difference runs the other way: generic models already knew far more about law than they know about our largely unpublished craft. **The model was never the product. The product is the warrant: defensible method, evidence and accountability.**

The honest economics: costs, returns and break-even

07

Estimated figures for a thirty-person independent strategic agency in 2026. Treat them as the starting point for your own spreadsheet, not the end of it: we would rather they were roughly right than precisely wrong.

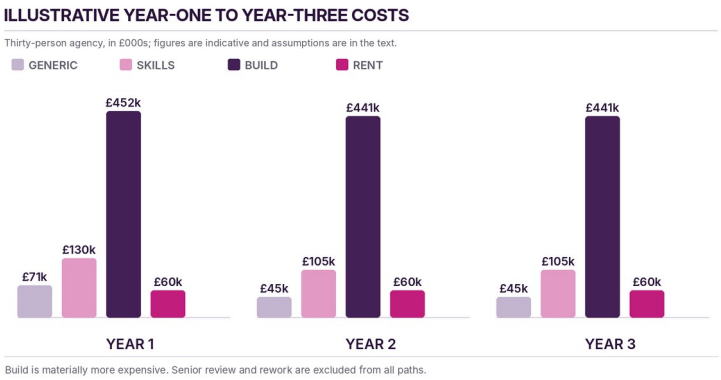
The agency we are modelling: thirty people, independent, strategic

Gross billings of £4.0m, less £1.2m of third-party pass-through, leaves **£2.8m of net fee income**: 58% payroll, 25% overhead, 17% operating profit, in line with agency benchmarks. Twenty-four of the thirty are in delivery and 60% of their time is production, so **roughly £780k of payroll is production**. Work in fee income, not turnover: AI compresses labour, not fieldwork.

Cost line	Generic	Skills layer	Build	Rent
Initial build cost	—	£60k	£370k+	—
Annual licence fees	£19k	£19k	£30k	£60k
Annual maintenance and hosting	—	£15k	£200k	Vendor
Annual upgrades and improvement	—	£45k	£185k	Vendor
Certification (SOC 2 and ISO 27001), year one	£33–70k	£33–70k	£33–70k	Vendor
Certification, annually thereafter	£19–33k	£19–33k	£19–33k	Vendor
Year one total	**£52–89k**	**£112–149k**	**£433–470k**	**£60k**
Annual cost thereafter	**£38–52k**	**£98–112k**	**£434–448k**	**£60k**

Source: Insightology internal cost data together with published estimates from a range of sources; 2026 midpoints. Certification: 2026 price guides for SOC 2 and ISO 27001 together. Skills layer assumes one full-time person building and maintaining it. Build is a floor: our own platform cost close to £900k in its first year. Senior checking and rework are excluded because they are time, not invoice; on the generic and skills paths they are usually the largest number on the page.

Only the build changes the order of magnitude; the other three sit in the low hundreds of thousands or less. **The skills layer, done properly, costs more to run than renting:** about £105k a year against £60k, once the person who maintains it, the rebuild cycle and certification are counted. And **the cheapest column is not the cheapest option:** a generic assistant looks unbeatable until you price the rework, every client-facing output lifted to standard by the people whose time costs most.



Midpoints of the ranges above. Build at the floor of its range.

What each path actually wins you

Costs are only half the decision, because AI pays back through two streams. **The bigger is revenue:** more and better proposals, winning work you would have declined, better thinking on every project, capacity without headcount, sharper prices and thinking priced at its premium. **The second is cost:** a smaller production bill, cheaper lost pitches, a lower polishing tax where output is client-ready, and compliance carried by someone else. Cost savings are bounded by your production bill; revenue growth only by the market. And savings only become cash when the freed capacity is monetised, so the comparison below redeploys freed time into revenue rather than cutting heads.

The return, worked: three years

Assumptions, stated so you can swap in your own. **Production saving:** 30% on rent and build, 15% on skills, well below reported task-level savings, because not every hour converts and a home-grown layer is used by fewer people. **Split:** a third to clients as sharper prices, a third through natural attrition, a third kept as margin; on £780k of production payroll that is £78k per third on rent and build, £39k on skills. **Revenue growth:** 10% on rent and build, 5% on skills, at a 50% margin: £140k and £70k a year. **Costs:** as above; build at the floor, about £450k in year one and £440k a year once live. **Timing:** benefits half-weighted in year one; a build earns nothing until year two.

ILLUSTRATIVE THREE-YEAR OUTCOMES

RENT A SPECIALIST PLATFORM	CREATE A SKILLS LAYER	BUILD IN-HOUSE
+£560k	**+£30k**	**−£740k**
NET OVER THREE YEARS	NET OVER THREE YEARS	NET OVER THREE YEARS
THREE-YEAR COST ~£180k	**THREE-YEAR COST ~£340k**	**THREE-YEAR COST ~£1.33m**
ANNUAL BENEFIT ONCE RUNNING	ANNUAL BENEFIT ONCE RUNNING	ANNUAL BENEFIT ONCE RUNNING
~£296k (£140k margin on new fee income + £78k attrition + £78k project margin)	~£148k (£70k margin on new fee income + £39k attrition + £39k project margin)	~£296k, but only from year two
BREAK-EVEN Within six months	**BREAK-EVEN During year three**	**BREAK-EVEN Not within three years**

The generic-assistant route is not shown: its cash cost is the smallest and its benefit the hardest to measure, because the saving is real at task level and largely consumed by checking.

Read it for shape, not decimals. **Revenue is the largest single component on every path. A build does not pay back at this size:** around £1.33m of cost against £590k of benefit over three years, at the floor of the cost range. **The skills layer barely washes its face:** around £340k of cost against £370k of benefit, ending about £30k up. That is what the completion rates in Section 4 look like on a timesheet. Now run it on your own numbers and real quotes, and be realistic: unless you are an experienced tech company, your in-house build is unlikely to massively outperform the rented version. Not impossible. Just unlikely.

The best alternative test and the moving target

08

Two rules sit behind the numbers you have just seen, and between them they settle most build-versus-rent debates before anyone opens a spreadsheet.

The best alternative test, with the maths

An AI capability wins you work only if it is better than what your competitors can get their hands on, and your competitors can rent the best specialist platform on the market tomorrow morning. So the return on building in-house is measured against that platform, not against doing nothing.

Put numbers on it. A rented platform costs £60k a year and delivers about £296k of annual benefit: a net gain of around £235k. A build carrying about £435k a year of running and keep-pace costs must generate around £670k of margin every year to beat that, and that is only to draw level, before it has recovered the £370k or more spent getting there. It has to do it by being meaningfully better than the rented platform, in ways clients notice and will pay for, sustained over years.

The moving target

A funded vendor does not stand still. Specialist platforms in this category invest in the region of £750k to £1.5m a year improving what they sell, and the foundation models underneath improve on their own cycle. So a build budget buys parity with today. Keeping parity means matching vendor-scale spend indefinitely, and surpassing it means out-thinking a company whose only job is the platform, and which does not also have to win and deliver research projects. The honest conclusion: **a build earns its keep only on things a vendor cannot replicate.** Proprietary data. Genuinely unique methods you would never share. A product you intend to sell in its own right.

Why a self-build needs a much higher return

A build carries materially more risk than renting: a large upfront commitment, delivery delay, performance uncertainty, key-person exposure and obsolescence. It should realistically target five to ten times the return of renting before it gets a green light. If renting nets around £235k a year, a build should be aiming at £1.2m to £2.4m a year. For an agency with £2.8m of fee income, that is not a bar you clear quietly. If you have a cost of capital and a view on execution risk, use those instead; the multiple is a stand-in for people who don't.

The cash flow problem

Even where the returns case survives, the cash flow rarely does. A build puts well over £370k out of the door in year one, before a single project benefits, and around £440k a year after that whether you win work or not. That is cash that would otherwise be senior hires, marketing, or the balance-sheet cushion every agency needs in a lumpy market, committed on a forecast.

A platform charging per won project has the opposite shape: nothing significant upfront, and the cost lands after the revenue it supports. A per-seat model sits between the two, a fixed cost that grows with headcount. One path bets the cash reserve on a plan; the other pays out of money already won.

If you take one thing from this section: never evaluate a build, or a skills layer, against having nothing. Evaluate it against the best thing you could rent in the year you would actually finish building, and demand five to ten times the return from a build, because that is what the extra risk and the locked-up cash cost.

Four routes: strengths and trade-offs

09

This scorecard uses the same eight assessment dimensions and dot ratings as **Choosing Your AI Technology Route**. It compares managed, business-grade AI configurations for a strategic market research agency, rather than consumer tools or individual products.

ASSESSMENT DIMENSION	GENERIC	SKILLS	BUILD	RENT
Fit with the agency's work	●●	●●	●●●	●●●
Output quality and research rigour	●	●●	●●●	●●●
Human control and accountability	●	●●	●●●	●●●
Security and data governance	●●	●●	●●●	●●●
Agency knowledge and accumulation	●	●●	●●●	●●●
Workflow and team operation	●●	●●	●●●	●●●
Implementation and ongoing ownership	●●●	●●	●	●●●
Economics and strategic freedom	●●	●●	●●	●●

● Limited / high burden ●● Mixed / material trade-off ●●● Strong / low burden

How the ratings work. More dots always means a more favourable position for the agency. Each row is the rounded average of four detailed criteria (32 in total). The scores are editorial judgements of defined configurations, not controlled product benchmarks or an overall league table.

What to take from the comparison

- **Specialist platforms lead on structured delivery.** Under these assumptions, both a successful in-house build and a suitable rented platform provide stronger research rigour, control and accumulated knowledge. Building buys customisation; renting reduces the engineering burden.
- **Generic and skills routes remain useful.** A managed generic assistant can be deployed rapidly for bounded tasks. A well-maintained skills layer adds reusable agency practice, but still relies on senior method authorship, upkeep and expert checking.
- **Economics do not produce a universal winner.** Licence fees, internal time, rework, project volume and supplier terms can change the answer. A high dot score cannot compensate for a failed data requirement, an unsupported method or unacceptable exit terms.

Where this leaves you

10

If you've read this far you likely know your answer, or you know which ninety-minute conversation your leadership team needs to have. The first thirty days are the same whichever path you take. **Count your production hours** on your last ten projects, split into production and thinking. **Find the AI already in the building:** who is using what on client data, under what rules. **Check your audit position** against the last three vendor assessments a client sent you. And **use the four-route scorecard** to agree the priorities with your leadership team, and write the decision down.

Then check any build instinct against the best alternative test. If the path is a platform, ask vendors to demonstrate the methodology, security and audit trail on your own work. Whichever path you land on, judge it against one standard: does it help you sell more warranted thinking, work you can prove, from thinking only your best people could do?

On the longer paths, don't stand still while you build. Reprice one service line to charge for outputs rather than days, write the one page of governance, and run a generic assistant properly in the meantime: on an enterprise tier, never on personal accounts for client work.

Some of the areas this paper touches get their own deep dive elsewhere: other white papers in this series cover why selling warranted thinking is how the industry will thrive in the AI era, a new pricing model for research after AI, what agencies can learn from the legal profession, and how to combine human powers with AI powers. And if the scorecard points you to a specialist platform, the fairest next step is a structured pilot on live projects, over a quarter, measured against the way you work today.

About Insightology

Insightology.ai is a specialist AI platform built for market research agencies, rather than generic AI with research prompts bolted on. It is method agnostic (qual, quant and mixed-method in one platform) and built for the complex, multi-market, multi-stage programmes strategic agencies actually sell. It covers the whole strategic project: decoding the brief, methodology design, proposal writing, survey and questionnaire design, discussion guides, fieldwork materials, analysis and reporting. Everything runs as a validated workflow with an audit trail, built to reproduce the quality of expert researchers consistently.

Agencies using it report time savings of up to 50% across proposals, fieldwork materials and analysis. Pricing is per won project with unlimited users. Proposals are free and unlimited, and credits only apply once a project is won and reaches fieldwork, so your cost follows your revenue rather than your headcount.

The argument in this paper applies to agencies doing strategic work, for clients who audit them, where warranted thinking is the real product. If that describes you, the quickest way to test any platform, ours included, is a structured pilot on real projects, measured over a quarter.

Go to www.insightology.ai to find out more and book a demo if it looks relevant to you

Appendix: the evidence in full

A

This paper makes three claims that rest on evidence beyond our own testing. **First,** a generic assistant's first draft is not client-ready, and checking and rewriting it eats most of the time it saved. **Second,** a skills layer on a generic assistant narrows that gap without closing it. **Third,** the professions that sell judgement under uncertainty have, having tried both, settled on specialist platforms, whose advantage lies in grounded, checkable work rather than a cleverer model.

For each study: what was measured, what was found and what we think it shows, including the parts awkward for our argument. Two cautions. Scores move, so hold on to the pattern rather than the numbers. And none of this is market research data: it is what adjacent professions measured first, in fields the models know far better than ours.

The studies behind the findings

Study	What was measured	What it found
OpenAI system card, 2025	How often its own reasoning models fabricate on recall questions	o3 invented answers to a third of questions about public figures and half of short factual questions
Stanford, 2024–25	Fabrication in leading AI legal research tools and a general model	Specialist tools 17% and 33%, general model 43%: roughly halved, not eliminated
METR randomised trial, 2025	Experienced developers in codebases they knew well, with and without generic AI	19% slower with AI, while believing they were 20% faster
BetterUp Labs and Stanford, 2025	AI-generated work that looks finished but is not (1,150 office workers)	40% had received it in the past month; nearly two hours to fix each piece
Upwork Research Institute, 2024	Effect of AI tools on workload (2,500 workers, four countries)	77% said workload rose; the top reason was reviewing and correcting output
Workday, 2026	Time saved with AI against time lost to rework	Nearly 40% of the time saved is lost again to rework
Minnesota and Michigan RCT, 2025	Law students with a specialist tool, a general model or no AI	Both tools far faster; 11 fabricated citations with the general model, 3 with the specialist
Harvey Legal Agent Benchmark, run by Vals AI, 2026	Leading general models with tools and document skills, on finished legal work judged by lawyers	The best completed about one task in four; the familiar assistants one in ten or fewer
Thomson Reuters CoCounsel, 2025–26	A specialist platform against practising lawyers, externally evaluated	73–90% across four task types, each more than ten points above the lawyer baseline
Vals AI Finance Agent Benchmark, 2026	Finished analyst work across several filings	Best model 61% with partial credit, 51% when every criterion had to be met
McKinsey, NBER, MIT NANDA, 2025–26	Whether AI programmes pay back	89% of around 6,000 executives saw no measurable productivity impact

HOW OFTEN GENERIC AI INVENTS THINGS

The model makers' own tests (OpenAI, 2025)

On questions about public figures, OpenAI's o3 fabricated answers to a third; on short factual questions, to half, and o4-mini to 79%, more than the models they replaced. Summarising a document it has been given, a leading model invents a few per cent at most.

What it shows. Invention is not a fault the next release will fix. It is worst where a model works from what it has read rather than what it has been given: material never published, like ours.

Specialist tools invent less, and not nothing (Stanford, 2024–25)

A Stanford team tested leading AI legal research tools and a general model on legal research questions. Fabrications and misstatements: 17% for Lexis+ AI, 33% for Westlaw's AI-Assisted Research, 43% for GPT-4.

What it shows. Building a profession's sources into the tool roughly halves the invention rate. It does not make output trustworthy on its own, which is why the warrant is people, method and a trail, not only software.

THE POLISHING TAX, MEASURED

Experts on familiar ground work slower with AI (METR, 2025)

A randomised controlled trial: 16 experienced open-source developers completed 246 real tasks in codebases they knew intimately, with and without generic AI tools. With AI they took 19% longer, and believed afterwards it had made them about 20% faster. Fewer than 44% of the AI's contributions were accepted without modification.

What it shows. The closest published analogue to a senior researcher with a generic tool or a skills layer. The time saved drafting is real; the time spent checking is larger, and invisible to the person spending it.

Almost-finished work shifts the burden downstream (BetterUp Labs and Stanford, 2025)

A survey of 1,150 American office workers, published in Harvard Business Review: 40% had received AI-generated work in the previous month that looked finished and was not, and each piece took nearly two hours to put right.

What it shows. The polishing tax is paid by the most senior person in the chain, because they are the one who can tell the output is wrong. In an agency, that is a director.

Workload rises; saved time leaks back (Upwork, 2024; Workday, 2026)

Upwork surveyed 2,500 workers in four countries: 77% said AI tools had added to their workload, most often through reviewing and correcting output. Workday found 85% of employees saving one to seven hours a week with AI, and nearly 40% of that time lost again to rework, more so early in adoption.

What it shows. A general tool on an unchanged process adds a checking job on top of the existing one. The 40% is the measurement behind the polishing tax in our cost model.

WHAT LAW HAS MEASURED

The randomised trial: specialist, general or nothing (Minnesota and Michigan, 2025)

Around 130 law students worked six realistic tasks with a specialist platform, a general reasoning model or no AI. Both tools roughly doubled productivity on several tasks and improved quality, the general model slightly more. But work with the general model contained eleven fabricated citations, and with the specialist three, fewer than with no AI. The authors call the counts descriptive, as the events were rare.

What it shows. On short, well-documented tasks a general model is as fast and as good as a specialist one. The specialist's advantage is reliability, and reliability is what the client is paying for.

Specialist beats raw model, then the raw models catch up (Harvey, 2024–26)

On Harvey's 2024 BigLaw Bench, Harvey scored 74% against roughly 60% for the raw models. Within eighteen months the base models reached around 90%, and Harvey's case for its product moved to the workflow, sources and agents around whichever model is best.

What it shows. Any advantage that lives in the model is on an eighteen-month clock. The comparator was the bare model; the next study supplies the skills layer.

A skills layer on a frontier model, measured on finished work (Harvey's Legal Agent Benchmark, run by Vals AI, 2026)

Each system gets a client's matter file and instructions and must produce a document a partner could review; a task passes only if every lawyer-written criterion is met. Vals AI, an independent evaluator, runs the leading general models with file, search and shell tools and Word, PowerPoint and Excel skills: a generic skills layer, not one written by lawyers. On 7 October 2026 the best agent completed about one task in four, and the frontier assistants most firms use one in ten or fewer. Criterion-level scores were far higher: the models met many requirements, then failed to finish the piece. Harvey has not published its own score on the same set.

What it shows. The nearest thing yet to an independent measurement of a skills layer on finished professional work. Three deliverables in four came back unfinished at best, and in an agency a senior finishes each one. One benchmark, in the profession the models know best; not a failure rate for market research.

Specialist platforms against practising lawyers (Thomson Reuters, 2025–26)

On the Vals AI legal research benchmark, CoCounsel scored between 73% and 90% across four kinds of legal task, each more than ten points above a practising-lawyer baseline. On fifty complex tasks of around six lawyer-hours each, it finished in under eight minutes, and experienced attorneys judged its answer stronger than the expert reference on nearly 40%.

What it shows. A specialist platform, independently evaluated, beating the professional baseline on production work; the vendor chose which tasks to publish. When Thomson Reuters later trained its own model and found it merely level with the frontier on reasoning, it deployed it first in structured document review: grounded, checkable production, where the warrant lives.

TAX, FINANCE AND ACCOUNTANCY

Research-grade questions and finished analyst work (Vals AI, 2026)

On research-grade tax questions requiring cited authority, the best general models score in the high seventies. On finished analyst work across several filings, the best scored about 61% with partial credit and 51% when every criterion had to be met. Mercor found frontier models faster and more accurate than junior accountants on well-defined tasks.

What it shows. The general models have conquered the well-defined question and the production end of the work. Nobody has shown them conquering the finished deliverable, or judgement.

WHY MOST AI PROGRAMMES SHOW NO RETURN

Adoption without redesign (McKinsey, NBER, MIT NANDA and BCG, 2025–26)

McKinsey: 78% of organisations use AI, about 21% have redesigned a workflow around it. NBER: 89% of around 6,000 executives see no measurable productivity impact. MIT NANDA (300+ initiatives): 95% of pilots show no measurable return, and bought tools reach production about twice as often as internal builds. BCG (nearly 12,000 workers): redesigning workflows makes measurable improvement 24 points more likely.

What it shows. A general tool bolted onto an unchanged process is the normal AI programme, and it returns nothing. That is the case for an enforced workflow and for change management.

What nobody has measured. There is no published, controlled comparison of a specialist platform against a well-built skills layer on the same tasks. Until someone runs that test, independently and on real deliverables, the defensible statement is the one this paper makes. If you would like to help run it for market research, we would welcome the conversation.

Sources

1. Marketing Research Institute International, State of the Market Research Industry, April 2026 (n=498). OpenAI, o3 and o4-mini System Card, April 2025; Vectara hallucination leaderboard.
2. Magesh, Surani, Dahl, Suzgun, Manning and Ho, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools", Journal of Empirical Legal Studies 22 (2025).
3. Becker, Rush, Barnes and Rein, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity", METR, July 2025.
4. Niederhoffer et al. (BetterUp Labs and Stanford), "AI-Generated 'Workslop' Is Destroying Productivity", Harvard Business Review, September 2025.
5. Upwork Research Institute, "From Burnout to Balance", July 2024; Workday, Beyond Productivity: Measuring the Real Value of AI, 2026.
6. Schwarcz, Manning, Prescott, Barry, Cleveland and Rich, "AI-Powered Lawyering", SSRN March 2025; Journal of Law & Empirical Analysis, 2026.
7. Harvey, "Introducing BigLaw Bench", 2024, and 2026 updates; Harvey, Legal Agent Benchmark, 2026, with the Vals AI held-out leaderboard as at 7 October 2026.
8. Thomson Reuters, CoCounsel evaluations, 2025–26, and Thomson-1-Large technical report, 2026 (LawSites, August 2026).
9. Vals AI, Tax Agent Bench, accessed 30 September 2026, and Finance Agent Benchmark v2, 2026. Mercor, "Human Baselines for Benchmarks", 2026.
10. McKinsey & Company, The State of AI, March 2025; National Bureau of Economic Research, Firm Data on AI, Working Paper 34836, 2026; Challapally, Pease, Raskar and Chari, The GenAI Divide: State of AI in Business 2025, MIT NANDA, July 2025; BCG, AI at Work 2026, June 2026.
11. Harvey customer figures: company announcement, 25 March 2026. Claude for Legal: launched 12 May 2026, as reported by LawSites, Artificial Lawyer and Fortune.
12. Costs: Insightology internal cost data and published estimates, for a thirty-person agency in 2026; agency shape from published benchmarks and the Agency Management Institute's 55:25:20 ratio; certification from 2026 Vanta, Drata, Secureframe and audit-firm price guides. Comparisons in Sections 3 and 5 are Insightology's own testing, offered as experience, not an independent benchmark.