



insightology.ai

Below the Surface

Finding true insight in market research data with AI, and proving it is true

A working guide to what generic AI and specialist analysers can find below the surface of your data, and how to check they are telling the truth.

A WHITE PAPER BY INSIGHTOLOGY

— OCTOBER 2026

QUAL

TENSION

Says price is irrelevant

...then explains every switch by price. Flagged in the respondent's own words.

EVIDENCE · VERBATIMS VERIFIED



QUANT

HIDDEN SPLIT

One flat score, two stories

One group delighted, another quietly leaving, cancelled out in the total.

EVIDENCE · TRACED TO SOURCE ROWS



OPEN-ENDS

ABSENCE

The leader nobody names

No unprompted mentions of the market leader. The silence is data.

EVIDENCE · THIN, FLAGGED



Add a source

Executive summary

01

Every AI tool can now summarise a data set in seconds. Few can analyse one, and fewer still can show you why you should believe what they found. That gap is where the value in research now sits.

A summary tells a client what their data says on average. They already know that: it is on their dashboards. What they pay an agency for is the reading below the surface: the tension, the absence, the single remark that explains the pattern.

The richest insight often sits where qual and quant meet, and that is where analysis is weakest. The two are almost always analysed apart, with qual used to decorate the quant, and rarely properly triangulated.

Generic AI is genuinely good at speed, first passes and sparring. It is structurally weak at the deep work. It is pulled towards the most likely reading, reads long material unevenly and gives different answers on different runs. When it is wrong, it is wrong silently, in the same confident voice.

Specialist analysers are built differently. The method lives in the software: every record read, sources joined, purpose-built passes hunting for specific kinds of hidden clue, and every finding traced to its source. They are not perfect, and they cost more than the tool already on your desk. **What they change is that errors become findable.**

SIX TAKEAWAYS FOR A RESEARCH DIRECTOR

- 1** Judge any AI tool by how far below the surface it reliably goes, not by how finished its output looks.
- 2** Where you have both, insist on qual and quant analysed together. That is where the deepest insights sit.
- 3** Use generic AI for exploration, sparring and drafting. Do not use it as the analysis of record.
- 4** Do not ask for zero hallucination. Ask for every claim to be checkable and every gap to be visible.
- 5** Test before you buy. Join qual and quant, plant a contradiction, run the same data twice and check every quote.
- 6** Keep the leap human. The machine marshals the evidence; your researchers decide what it means.

The surface of good research can now be generated on demand, so the surface has stopped carrying value.

The summary is not the insight

02

Clients are not short of information. Their rivals have the same dashboards, the same trackers and the same reports. A widely cited Gartner estimate puts unstructured data at 80–90% of all new enterprise data, growing three times faster than the structured kind.

What is scarce is a reading that tells them something true nobody else has seen. That is insight, and it is worth separating from its two cheaper cousins. A summary compresses what the data says. A finding states a fact the data shows. An insight explains what is actually going on, and changes what the client does next.

THREE READINGS OF THE SAME DATA

Summary

WHAT THE DATA SAYS

Customers are broadly happy but want faster support.

Finding

A FACT THE DATA SHOWS

Younger customers rate support far worse than older ones.

Insight

WHAT IS ACTUALLY GOING ON

For younger customers it is not about speed: they want great self-serve support tools, not a human or a chatbot.

Illustrative example.

Insight almost never sits in the average. This is **the Tyranny of the Average**: a summary reports the centre of the data, and the value is at the edges. The more efficiently you summarise, the more completely you erase the edges.

The problem predates AI. Researchers have always sampled their own evidence: twelve of sixty transcripts read properly, the qual in one deck and the quant in another, last year's study in a folder nobody opens. Forsta's widely quoted benchmark puts 80% of a researcher's time into preparing and organising data, leaving 20% for analysis and insight.

AI changes this in two directions at once. It can read everything, which no team on a deadline can. And it can make a surface summary look so finished that nobody digs any further. **The first is the opportunity; the second is the new risk.**

Which one an agency gets depends on the tool and how it is built. Our companion paper on insight as the new currency makes the commercial case. This paper is about how the analysis is actually done, and how to tell the tools apart.

Where insights often hide: between qual and quant

03

Most research projects collect both qual and quant. Very few analyse them together. **That gap is where the richest insights go missing.**

How it is usually done

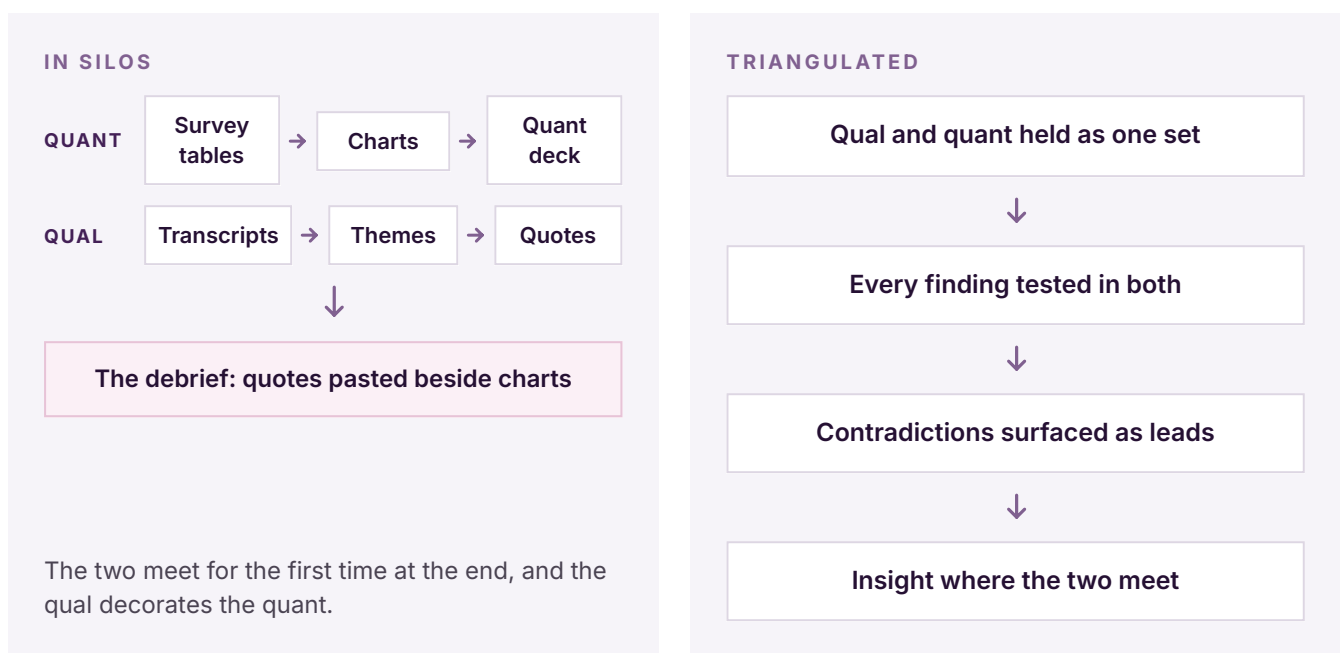
Qual and quant are commissioned together, then analysed apart: by different people, with different tools, on different timelines, into different decks. The two often meet for the first time in the debrief.

Qual as decoration

When they do meet, qual is usually there to decorate the quant. A verbatim is dropped beside a chart to bring alive a number that was already decided. Triangulation is promised in the proposal and rarely done in the analysis.

Why it happens

Not through lack of will. Nobody on a deadline can hold sixty transcripts and four hundred respondents' answers in their head at once, and the tools keep them apart: one for tables, another for transcripts.



This does not mean a specialist analyser only earns its keep on mixed studies. On qual-only or quant-only projects it still does better than generic AI: every record read, one method applied every time, every finding traced. Having both simply adds depth.

The richest insight often sits where qual and quant meet, and that is where analysis is weakest.

03 · WHERE INSIGHTS OFTEN HIDE

Four places the insight hides

Real triangulation means testing every finding in both sources. Four kinds of insight only appear when you do:

The why behind a number

The survey shows what is happening; the interviews explain why. Without the why, a number is a fact the client cannot act on.

The number that tests a story

A compelling theme from the groups may describe three people or three thousand. The survey says which.

The contradiction

People say one thing in the groups and choose another in the survey. The gap between them is often the most valuable finding in the project.

The segment the survey finds and the qual explains

The quant spots a group behaving differently; the qual tells you who they are and what drives them.

The example, traced to its sources

Take the support example from Section 2. Each step comes from a different source.

QUANT

The finding

Younger customers rate support far worse than older ones.



QUAL

The insight

For younger customers it is not about speed: they want great self-serve support tools, not a human or a chatbot.



QUANT

The sizing

How many would rather fix problems themselves, and in which segments?

Neither source finds that insight alone.

Illustrative example.

A tool that cannot hold qual and quant together cannot find the insights that live between them.

Using generic AI for analysis: what it does well

04

Underrating generic AI is as expensive a mistake as overrating it. So before the critique, the credit. Six things it does genuinely well:

Speed

A first pass of themes from a set of transcripts in minutes, not days. For a researcher getting their bearings, that is a real gift.

Sparring

Ask it for three alternative readings of a finding, or to argue against your favourite one. It never tires and has no ego in the answer.

Small, single-source jobs

A handful of interviews, or a few hundred open-ends that fit comfortably in one conversation, are well within its reach.

Drafting and editing

Once the analysis is done and checked, it turns notes into clean prose, headlines and slide copy quickly.

Cost and access

No procurement, no implementation, and it is probably already on every desk in the agency.

Direction

A skilled researcher can push it part of the way below the surface with the right prompts: ask for contradictions, for what is missing, for the minority view.

On short, well-defined tasks the general models are now strong and converging. The trouble starts when the task is long, mixed and has no single documented answer. That describes almost every piece of strategic analysis.

51%

Best general model, every criterion met

The evidence shows the same shape. On a 2026 independent test of finished analyst work, connecting data across several documents with an insight expected at the end, the best general model met every criterion about half the time.

Source: Vals AI, Finance Agent benchmark, 2026.

Generic AI: where it breaks below the surface

05

None of what follows is a bug the next model release will patch. Each limit comes from how generic assistants are built, and how they are used in practice.

01 Pulled to the average

A model trained to produce the most likely continuation is pulled towards the most familiar reading. A single remark that outranks a pattern across four hundred is exactly what it is built to smooth away.

02 Long material read unevenly

Models do not attend evenly to everything they are given. Material in the middle of a long input is used less reliably than material at the start or end (Liu et al., 2024). In a sixty-transcript project, the middle is most of the data.

03 Batches, not a set

In practice the material goes in batches, and each batch is analysed on its own terms. The coding frame drifts unless someone re-pastes it every time. Transcripts go into one chat and cross-tabs into another, so qual and quant never meet.

04 Not repeatable

Ask the same question of the same data twice and you can get two different theme structures. A finding that appears on one run and not the next should not go in front of a client.

05 A manner, not a method

The models have read very little of how strategic research is actually analysed, because that craft was rarely written down. We call this **the Blank Slate Problem**. A prompt can direct a model; it cannot educate it.

06 Numbers generated, not computed

Unless the tool is running code, a figure in a chat answer is predicted text, not a calculation.

07 No trail, no memory

A chat log is not an audit trail, and a quote in a summary cannot reliably be traced to the person who said it. Each session starts close to zero, so the agency's own conventions and past findings are never in the room.

05 · GENERIC AI: WHERE IT BREAKS BELOW THE SURFACE

The cost: the Polishing Tax

A chat summary of sixty transcripts looks finished and is not. So a senior person spends the afternoon checking it against the transcripts anyway. That is **the Polishing Tax**, and it falls on the most expensive person on the project.

HOW THE TIME MOVES



It has been measured. Three studies show the same pattern from different angles.



Will a better model fix it?

Partly. Models will keep improving at short tasks. But most of the limits above sit in the system around the model: how material is fed in, whether a method is applied, whether results are checked and traced.

A better engine in the same chat window gives you a faster version of the same problem.

Hallucination in research analysis

06

The industry calls it hallucination. In research terms it is simpler: **any claim the evidence does not support, stated as if it does.**

It is not a malfunction. A model built to produce plausible text produces plausible text. Most of the time that is also correct; on thin ground it is not, and there is no light on the dashboard to tell you which you got.

How often it happens

Far more often than most people assume, especially when a model works from memory rather than from material you have given it. In OpenAI's own 2025 tests of its o3 model, it fabricated answers to a third of questions about public figures and to half of short factual questions.

1 in 3

answers about public figures fabricated by OpenAI's o3, working from memory
OpenAI system card, 2025

1 in 2

short factual questions answered with a fabrication by o3; 79% for the smaller o4-mini
OpenAI system card, 2025

Giving a model your fieldwork makes invention much less likely. It does not make it zero. And a few errors in a hundred sounds tolerable until you remember that you do not know which few.

Why it cannot be prompted away

The standard defences help. Tell the model to answer only from the material supplied. Ask for a quote or cell reference for every claim. Keep inputs focused, and run a second pass that checks each claim against its source.

Those defences reduce the risk. They do not remove the checking, and on a generic assistant they add steps of their own.

So every figure and every quote has to be checked.

06 · HALLUCINATION IN RESEARCH ANALYSIS

What it looks like in a debrief

In research work, the drift takes six recognisable forms:

Filled gaps

A plausible answer where the material has none, instead of "this isn't covered".

Wrong details

A figure transposed, a statistic pinned to the wrong segment or the wrong wave.

Invented quotes

Words put in a respondent's mouth that sound right and were never said, or a real quote given to the wrong person.

Blended sources

Your data mixed with the model's general knowledge, or details from two files merged into one.

Uneven reading

Material in the middle of a long transcript treated less reliably than the start or the end.

Overgeneralisation

"Three of twelve respondents said this" becomes "respondents generally felt this".

Below the surface, two more matter: **contradictions smoothed into one tidy theme, and connections asserted that the data does not make.**

Why it matters more below the surface

Surface findings are easy to check, because everyone can see the top line. Below-the-surface findings are the opposite. A tension between segments, a remark made once, an absence: these are the hardest to verify by eye and the most likely to change a client's decision.

So exactly where you most need confidence, a generic tool gives you least.

The right promise

Be wary of anyone who promises zero hallucination.

The right promise is different: every claim checkable, every gap visible, and errors findable in minutes rather than hours.

07

Specialist analysers: how they are built to look below the surface

A specialist analyser is not a generic model with a better prompt. **The method lives in the software, not in whoever happens to be typing.**

In plain English, “coded models” means the analysis follows a fixed, tested sequence of steps, each built for one job. The language model works inside that structure rather than being in charge of it. Five features matter most.

01 Qual, quant and both together

It analyses a set of interviews on its own, or a survey on its own, more reliably than a generic tool, and it can hold the two together. That adds depth: a pattern in the numbers is explained by what people said, and a claim in the interviews is tested against the survey.

02 Everything read, held as one set

Every transcript, open-end, data table, document and prior wave is loaded and indexed together. Nothing is sampled and nothing is analysed in isolation.

04 Numbers computed, not generated

Counts, percentages and cross-tabs are calculated from the data, then described in words.

03 A defined pipeline

Code, cross-reference, triangulate, then interpret. The same steps in the same order on project 1 and project 200, so quality stops depending on who picked up the job.

05 The agency's own knowledge

Coding frames, battery structures, conventions and past findings sit alongside the method. The analysis reflects how this agency works, and improves with every project.

For centuries we have used statistics to find hidden patterns in numbers. There was no equivalent for numbers and words together. That is what a well-built analyser is: **statistics for mixed data.**

07 · SPECIALIST ANALYSERS: HOW THEY ARE BUILT TO LOOK BELOW THE SURFACE

The clues they are trained to hunt for

Generic tools look across the data. Specialist analysers are built to look into it, with purpose-built passes trained on what experienced researchers look for. Six examples:



Where qual and quant meet, or disagree

Qual read against quant, this wave against the last, open-ends against closed questions, what people say in interviews against what they choose in the survey. These connections only exist between sources, so they only appear when the sources are held together.



Tensions and contradictions

Within one respondent, between what is said and what is done, between segments. A respondent who calls price irrelevant, then explains every switch by price, is telling you something.



The single voice that matters

The remark made once that outranks a pattern across four hundred. The pass flags it as rare and potentially significant, rather than averaging it away.



Absences

What is conspicuously not said, not chosen, not mentioned. If nobody in a category study names the market leader unprompted, that silence is data.



Hidden splits

Segments whose views cancel each other out in the total. A flat satisfaction score can hide one group delighted and another quietly leaving.



Early change

The drift that is visible before it becomes a trend: a new phrase creeping into open-ends, a small segment moving ahead of the rest.

None of these clues is an insight in itself. Each is a lead worth a researcher's attention, surfaced from the whole data set rather than the corner someone had time to read.

Specialist analysers: how they are built to expose hallucination

08

You cannot prompt your way to trustworthy analysis. You have to engineer it.

A well-built analyser treats every claim as unproven until it can point to its evidence, and builds that proof into the output. Seven controls do the work:

01 Grounding

Every finding is linked to the evidence behind it: the respondent, the row, the timestamp, the page. One click from claim to source.

02 Verbatim checking

Before a quote appears, code checks it word for word against the source. A quote that does not match exactly is corrected or removed, never paraphrased inside quotation marks.

03 Counts behind claims

"Respondents felt" is replaced by how many, out of how many, in which segment. Overgeneralisation becomes visible on the page.

04 Evidence flags

Thin evidence is labelled as thin. Where the material does not answer a question, the output says so rather than filling the gap.

05 Counter-evidence

The case against each finding is surfaced alongside it: the respondents who said the opposite, the segment where it does not hold.

06 Consistency

The same data through the same method gives the same answer. A finding that is unstable across runs is flagged, not presented.

07 Audit trail

Every run, step and intermediate output is logged. A director can see how an output was built before it reaches a client, and the client can follow the trail afterwards.

PLAUSIBLE BECOMES WARRANTED

Plausible means it looks like the real thing. Warranted means you can show why it is: the method, the data, who checked it and the trail back to a real respondent.

08 · SPECIALIST ANALYSERS: HOW THEY ARE BUILT TO EXPOSE HALLUCINATION

Checking becomes a spot-check

With these controls, checking becomes a spot-check rather than doing the analysis again. That is where the hours are actually returned, and where the Polishing Tax shrinks.

A GENERIC SUMMARY

Check everything

You cannot tell which figure or quote is wrong, so every one goes back to its source. In practice, the analysis is done twice.

A TRACED ANALYSIS

Spot-check

Every claim carries its source, its count and the case against it. A director follows the trail where it matters most.

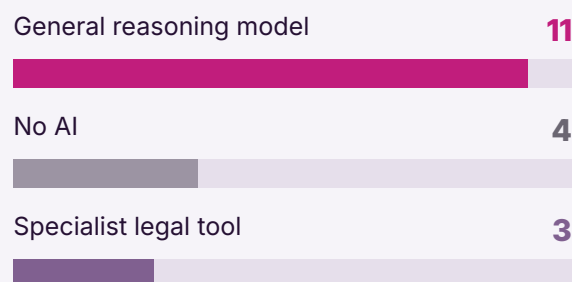
The honest claim

These controls expose and contain hallucination. They do not abolish it.

In a 2025 randomised trial, work done with a specialist legal tool contained three fabricated citations, against eleven with a general model and four with no AI (Schwarcz et al.).

The authors treat the counts as descriptive because the events were rare.

FABRICATED CITATIONS IN A RANDOMISED TRIAL



Schwarcz, Manning, Prescott, Barry, Cleveland and Rich, Minnesota and Michigan, 2025.

The point is not that specialist tools never err. It is that **their errors become findable**, quickly, by the person whose name goes on the debrief.

Specialist analysers: the honest drawbacks

09

Specialist tools are not a free lunch. A paper that only made the case for them would be a brochure. **Here is the other side.**

Cost and procurement

Another contract, another security review and another line on the P&L, set against a generic tool already on every desk. Model your year-three bill, not your pilot bill.

Less flexible

A defined pipeline is excellent at the analysis it was built for and less good at the odd one-off question. For those, a generic assistant is often quicker.

Slower than a chat reply

A multi-pass analysis with checks takes minutes or hours, not seconds. That is the price of reading everything and checking it.

The method is the vendor's

You are buying methodology in software. If the vendor cannot show you the method, who validated it and how, it is a black box with better branding.

Lock-in

Ask what happens to your project record, coding frames and outputs if you leave, and in what format you get them back.

The label proves nothing

Many tools sold as specialist are a thin wrapper on a generic model that summarises and visualises. Same limits, nicer interface. Only testing tells the difference.

Adoption takes work

Researchers used to working from a sample have to learn to work from the whole picture, and to trust a trail rather than their own re-reading. That is a change in habit, not just a new login.

Not infallible

Traceability makes errors findable, not impossible. Senior review stays, even if it gets much shorter.

Choosing and testing: side by side

10

Most agencies will use both kinds of tool. **The real question is which one carries the analysis of record.**

	GENERIC AI	SPECIALIST ANALYSER
Qual and quant together	Only if a person stitches the batches together	Qual, quant and open-ends joined by design
Coverage	Batches; the middle of long inputs read unevenly	Every record read and held as one set
Below-the-surface reach	Pulled to the average; depends on the prompter	Purpose-built passes for tensions, absences, splits and outliers
Repeatability	Two runs, two theme structures	Same method, same answer; unstable findings flagged
Quote integrity	Can invent, blend or misattribute	Quotes checked word for word against the source
Numbers	Predicted text unless code is run	Computed from the data
Hallucination	Fails silently	Thin evidence flagged; counter-evidence shown
Audit trail	A chat log	Finding-to-source trail and run history
Agency knowledge	Starts close to zero each session	Frames, conventions and past findings carried forward
Flexibility	Very high	Lower outside the analysis it was built for
Speed of reply	Seconds	Minutes to hours for a full analysis
Cost and set-up	Low; already on desks	Contract, onboarding and security review
Best use	Exploration, sparring, drafting	The analysis that goes to the client

10 · CHOOSING AND TESTING: SIDE BY SIDE

A sensible division of labour

USE GENERIC AI TO

Get your bearings, test framings and draft.

USE A SPECIALIST ANALYSER FOR

The analysis that goes in front of a client with your name on it.

A buyer's test you can run in an afternoon

Take a past project where you already know the answer, and put any tool through these steps:

01 Join qual and quant

Give it interviews and survey data from the same study, ideally one where the two point in different directions. Does it connect them and spot the contradiction, or report them side by side?

05 Check every quote

Search for each verbatim in the source. Any that does not match exactly is a fail.

02 Run it twice

Same data, same instructions. Do the themes match?

06 Ask for the count

For each theme: how many respondents, out of how many, in which segments?

03 Plant a needle

Add one striking remark to one transcript. Does the tool find and flag it, or average it away?

07 Ask what it could not see

A good tool tells you where the evidence is thin. A poor one has an answer for everything.

04 Plant a contradiction

Have one segment say the opposite of another. Does it report the split, or blend it?

08 Compare with the original debrief

What did it find that the team missed, and what did it miss that the team found?

For due diligence on the supplier itself, start with ESOMAR's 20 Questions to Help Buyers of AI-Based Services. It asks whether a supplier is responsible. **The test above asks whether the tool actually works.**

The human leap stays human



Nothing in this paper says machines find insights. They find clues, at a scale and consistency no team can match. With the right engineering, they show their working. **The leap from clue to insight stays human.**

PATTERN

The machine's territory

- Reads the whole data set
- Hunts for tensions, absences and outliers
- Shows its working, from claim to source
- Marshals the evidence for and against each reading

What changes is what the researcher leaps from. Not the twelve transcripts there was time to read, but the whole data set, with every tension, absence and outlier laid out and every claim traced to its source.

Pattern is the machine's territory. Resonance is ours: knowing which single thing matters for this client, this decision, this month. A model weighs evidence by how often it appears; a researcher weighs it by what it means.

RESONANCE

Ours

- Knows which single thing matters
- Weighs evidence by what it means
- Makes the leap from clue to insight
- Puts a name to it and defends it

The machine can marshal the evidence for and against each reading. Someone still has to choose, put their name to it and defend it in front of a finance director who is paid to be unconvinced. That is insightfulness and judgement, two of the seven human powers, and AI raises their value rather than threatening it.

This is what we mean by **warranted thinking**: strategic output you can prove. The warrant comes from the system: validated method, real respondents, governed AI and a trail to the source. The thinking comes from your people.

The surface of good research can now be generated on demand, so the surface has stopped carrying value. What clients will pay for is a reading that tells them something true nobody else has seen, dug out from everything they already hold, with the evidence to prove it.

About Insightology

12

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.

Its **Multi-Modal Analyser**, also available as a standalone service, reads transcripts, survey data, open-ends and documents (Word, PowerPoint, PDF or Excel, with SPSS coming soon) as one set, looks below the surface rather than summarising, and traces every finding to its source.

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 20-minute demo of the Multi-Modal Analyser on your own data

Sources

- 1 Andersson, T. (Forsta), "The 80/20 problem in market research", GreenBook and Research World, 2025.
- 2 Gartner estimate of unstructured data (80–90% of new enterprise data), as reproduced in ESOMAR, Global Market Research 2022.
- 3 Liu, N. F. et al., "Lost in the Middle: How Language Models Use Long Contexts", Transactions of the ACL 12 (2024).
- 4 OpenAI, o3 and o4-mini System Card, April 2025.
- 5 Schwarcz, Manning, Prescott, Barry, Cleveland and Rich, "AI-Powered Lawyering", Journal of Law & Empirical Analysis, 2026.
- 6 Workday, Beyond Productivity: Measuring the Real Value of AI, 2026.
- 7 Niederhoffer et al. (BetterUp Labs and Stanford), "AI-Generated 'Workslop' Is Destroying Productivity", Harvard Business Review, September 2025.
- 8 Vals AI, Harvey Legal Agent Benchmark leaderboard (October 2026) and Finance Agent Benchmark (2026).
- 9 ESOMAR, 20 Questions to Help Buyers of AI-Based Services, March 2024.
- 10 Jay, M., How to Build an AI-Enabled Market Research Agency (2026).