REPORT

The State of AI in Sales, *2026*

Ten sales functions graded against independent evidence only. What is proven, what is oversold, and the eleven places the evidence does not exist.

Independent evidence only · no vendor-authored sources

What the evidence actually supports, function by function

Published by Atrium · 2026 edition

Executive summary

In one line: adoption is near-universal, returns are not, and the functions with real evidence behind them are almost never the functions being bought.

The numbers that matter

Organisations using AI in at least one function	88% *(McKinsey, 1,719 respondents)*
Attributing any positive EBIT contribution to it	37% *(unchanged from 2025)*
Reporting tangible ROI	8% *(KPMG, 2,110 C-suite leaders)*
Achieving substantial ROI across the business	5% *(BCG)*

Sales orgs reporting a return of 50%+	25% *(Gartner, 210 chief sales officers)*
Sales orgs reporting a NEGATIVE return of 50%+	20% *(same survey)*
Sales & Marketing as a deployment function	#1, at 52% *(US Census, nationally representative)*
Sales & Marketing daily generative-AI usage	Lowest of any function, 34% *(Wharton/GBK)*
Quantified commercial-outcome claims that are vendor-authored	~100% *(this report's own search)*

The puzzle: sales attracts the most AI deployment and produces the least daily use and the flattest returns. One sales organisation in five is measurably worse off than before it started.

Where the value actually is

Ten sales functions, graded against independent evidence only. Ranked by strength of the case.

#	FUNCTION	VERDICT	BEST INDEPENDENT EVIDENCE
1	Responding to inbound / customer-initiated contact	Proven	**+16.3% on sales** among 44,614 consumers. Push messaging to 13.7 million returned +1.6%, not significant.
2	Onboarding and activation	Proven	**First-week churn halved**; field experiment, 366 customers treated of 2,673. The +46.6% usage figure is cumulative over eight months; the effect on activity decays within about a week.
3	Retention targeting by responsiveness	Promising	**≥4% firm profit at identical spend**, purely by changing who is targeted
4	Quote and pricing support (hybrid)	Promising	**+7.8% profit** vs +4.9% for full automation; 67,851 quotes
5	Churn prediction	Proven	Reliable to about 77% accuracy ceiling
6	Levelling up weaker performers	Proven pattern	**+34% novices**, +43% below-average; replicated across 4 studies

#	FUNCTION	VERDICT	BEST INDEPENDENT EVIDENCE
7	Call transcription *(time saved only)*	**Proven**	Moderate
8	CRM hygiene and data capture	**Promising**	a randomised trial of 238 physicians, as the closest proxy
9	AI qualification of inbound	**Promising**	randomised trial, 6,255 leads — but disclosure-fragile
10	Call coaching	**Promising**, conditional	Works only when feedback volume is restricted

And where the evidence is weak or absent:

FUNCTION	VERDICT	WHY
Outreach personalisation across the business	**Oversold**	Personalisation premium about **0.43 percentage points** (76,977 recipients)
Firm-initiated follow-up and nurture	**Oversold**	**No significant effect** on 13.7 million consumers
Sales forecasting	**Oversold**	No commercial product has ever been independently evaluated
Churn *prevention* as practised	**Oversold**	An at-risk intervention **raised** churn from 6% to 10% in a randomised trial
Monitoring as a performance lever	**Disproven**	Two meta-analyses covering 23,461 people found the effect on performance is **zero**
Prospect research, meeting prep, proposals, routing logic	**Unproven**	No independent efficacy evidence exists for any of them
Speed-to-lead	**Unproven**	The 21x and 100x multipliers trace to 2007/2011 studies funded by a company selling lead-response software. No independent study has tested response latency since.

The five things to take away

1. **Reactive beats proactive.** The single cleanest finding in the field, causally identified at

one company's seven experiments. Where the customer opened the conversation, AI helped. Where the firm opened it, it did not.

1. **AI redistributes performance, it does not lift everyone.** Value concentrates at the bottom of the skill distribution. Deployments justified as a uniform productivity multiplier are measured against the wrong expectation.

1. **Hybrids beat automation in every controlled comparison.** More autonomy is not more value.
2. **Time saved is not money saved.** The conversion step almost never gets built, and self-reported savings are unreliable by up to 39 percentage points.

1. **The evidence measures substitution.** Every study here ran inside an organisation that already staffed the function. **An Oversold grade means "does not beat a functioning team" — not "produces nothing."** For under-resourced teams the counterfactual is zero, and nobody has measured it. See *Who this evidence does not cover*, end of Part I.

How to read the grades

Proven — independent, controlled evidence of a commercial outcome. **Promising** — real evidence, but conditional, indirect, or from a single study. **Unproven** — searched for specifically; no independent evidence exists either way. **Oversold** — independent evidence exists and does not support the claims being made. **Disproven** — independent evidence shows the opposite.

About this report

Almost every number circulating about AI in sales was published by a company selling AI for sales. Across roughly sixty targeted searches conducted for this report, **essentially 100% of quantified commercial-outcome claims — win rates, reply rates, ramp times, conversion lift — were vendor-authored or vendor-derived.** Statistics appear with no study named, no sample size, and no date, and are then cited by the next article as established fact.

This report takes the opposite approach. Every figure below names its study, its sample size and its date. Nothing authored by a company selling the thing being assessed is

used as evidence. Where credible independent measurement does not exist, we say so and leave the gap visible, because a documented absence of evidence is more useful than a confident number that came from a brochure.

There are eleven such gaps in this report. They are listed at the end. The largest of them is a limitation of the entire evidence base, and it is described at the end of Part I.

Contents

Part I — The spending and the returns

Adoption is close to universal. Returns are not.

McKinsey, The State of AI: Global Survey 2026 (surveyed 1,719 respondents across 97 nations between 4 May and 8 June 2026): **88% of organisations now use AI in at least one business function.** But only **37% attribute any positive EBIT contribution to AI**, essentially unchanged from 2025, and just **6% qualify as high performers** — also unchanged. Adoption scaled. Impact did not.

KPMG Global AI Pulse, Q1 2026 (2,110 C-suite leaders across 20 countries): **8% report tangible ROI.** Average enterprise AI budget: **$186 million.** Only 7% had established ROI at all; 24% reported investor pressure to demonstrate it.

BCG 2026 AI Radar: 5% of enterprises achieve substantial ROI across the business.

IBM CEO Study: 25% report delivering expected ROI; **56% of CEOs say no significant benefit has materialised.**

And for sales specifically — Gartner, 210 chief sales officers and senior sales leaders, fielded in the first two months of 2026:

> *25% of sales organisations report a return of 50% or more on their AI investments. 20% report a negative return of 50% or more.*

One organisation in five is measurably worse off than before it started.

Sales is where the money goes, and where it is used least

US Census Bureau, Business Trends and Outlook Survey / CES Working Paper 26-25 — nationally representative, reference period November 2025 – January 2026: **18% of US firms used AI** (32% employment-weighted). Among adopting firms, **Sales and Marketing is the number one deployment function at 52%**, ahead of Strategy and Business Development (45%) and IT (41%). But **57% of AI-using firms use it in three or fewer functions**, only **2% report labour reductions**, and 66% use AI purely to augment existing work.

Set against that, **Wharton Human-AI Research with GBK Collective** (~800 US enterprise decision-makers, firms with 1,000+ employees and $50M+ revenue, surveyed June–July 2025) found that **Marketing and Sales has the lowest daily generative-AI usage of any function surveyed, at 34%** — behind IT (68%), Procurement (57%), Operations (43%) and Finance (39%).

The tension in one line: sales attracts the most AI deployment and produces the least daily use and the flattest returns. That is the puzzle this report exists to explain.

A note on the most-cited statistic in the field

The figure "95% of generative AI pilots deliver zero measurable P&L impact" comes from *The GenAI Divide: State of AI in Business 2025*, MIT NANDA, July 2025. It has been quoted in most coverage of this subject, and it moved AI-linked equities when *Fortune* publicised it in August 2025.

It should be treated with care. The study is a **desk review of 300+ publicly disclosed initiatives, 52 structured organisational interviews, and 153 senior leaders surveyed at four industry conferences** — a convenience sample. It is not peer-reviewed. NANDA is supported by Google and Anthropic, and the report advocates architectures associated with those vendors. It gives two different figures, 70% and 50%, for the same quantity in different passages. The 95% itself derives from a **pipeline-conversion statistic** for custom-built enterprise tools reaching production, not from measured P&L; general-purpose LLMs in the same report deployed at 40%.

Most importantly, the report disclaims its own use-case breakdowns:

> *"Sub-category and use-case breakdowns should be treated as directional at best. Subcategories reflect synthesized notes and anecdotal patterns, rather than precise accounting."*

We cite it here as an indicative signal of sentiment, and nowhere else. Everything that follows rests on stronger ground.

Who this evidence does not cover

Two limitations run through everything that follows, and both are large enough to change how the verdicts in Part II should be read.

Two limitations, and the second is the one that carries the report's organising principle.

One: the population the evidence was drawn from is mostly not B2B. The reactive-beats-proactive principle rests on Fang et al., run at a cross-border consumer retail platform. Luo's qualification result is consumer telemarketing for financial products. Retana's onboarding experiment is self-serve cloud infrastructure. Brynjolfsson's support agents handle consumer software support. The audience for this report is B2B, where the buying unit is a committee rather than a person, the consideration window runs months rather than minutes, and a single deal is worth more than a platform's entire experimental arm.

Nothing in the literature establishes that the pattern transfers. The mechanism is plausible in both settings — a buyer who has just raised their hand is more receptive

than one who has not — and the direction of the B2B evidence that does exist, on outbound personalisation and firm-initiated follow-up, is consistent with it. Treat the principle as a well-supported hypothesis carried across a population boundary, and weight it accordingly against your own data.

Two: every study in this report was conducted inside an organisation that already staffed the function being tested. Fang et al.'s firm-initiated null compares AI-driven outreach against an outreach operation that already existed. Brynjolfsson's +14% average and +34% for novices is measured across 5,179 people already employed as support agents. Karlinsky-Shichor & Netzer's pricing result is 17 representatives who were already producing quotes. Retana's onboarding experiment ran at a provider that already had an onboarding motion to improve.

This is not an oversight by the researchers. Measuring a lift requires a baseline, and a baseline requires someone to have been doing the work. It does, however, mean that **every verdict in Part II is a verdict about substitution** — whether AI outperforms the human currently doing the job — and not about whether the work is worth doing.

The population absent from this literature is the one where the function is not performed at all: the founder-led sales motion with no SDR, the team with no pricing desk, the company where follow-up happens when someone remembers. For that population the counterfactual is not a competent human. It is zero. **No study located for this report measures that case**, and the distinction matters commercially: an Oversold grade in Part II means "does not beat a functioning team," not "produces nothing."

The one body of evidence that speaks toward this population points the other way. The skill-distribution finding in Part IV — replicated across four independent studies in different domains — is that value concentrates where existing capability is lowest: **+34% for novices, +43% for below-average performers, $14.85 per quote for low-expertise reps versus $5.21 for high.** The consistent direction is that the less well the work is currently done, the more the intervention is worth.

Extending that line from *least capable person doing the job* to *nobody doing the job* is an extrapolation, and we mark it as one rather than pretending it is a measured result. But it is the direction the evidence leans, and an absence of measurement is not a measurement of absence. It is the eleventh and largest gap in this report.

Part II — Where the money is actually being made

Correction before anything else

The thesis I proposed in the outline rested on the MIT NANDA study. **On verification, that study cannot carry it, and neither the figures nor the methodology I quoted were right.**

What I said it was: 300 AI deployments, 150 leadership interviews, a 350-person employee survey.

What it actually is: a desk review of **300+ publicly disclosed initiatives**, **52** structured organisational interviews, and **153** senior leaders surveyed at four industry conferences. There is no employee survey. It is not peer-reviewed. It was distributed via a gated form rather than an MIT publication channel. NANDA is supported by Google and Anthropic, and the report advocates architectures associated with those vendors. It contradicts itself on the sales-and-marketing budget share, giving both 70% and 50% in different passages.

And the "95% of pilots deliver zero P&L impact" figure is not a measured P&L finding. It is a **re-framing of a pipeline-conversion statistic** for custom-built enterprise tools reaching production (60% evaluated → 20% piloted → 5% production). General-purpose LLMs in the same report deployed at 40%. Wharton's Kevin Werbach, reviewing it, found no further support for the claim.

Most damningly, the report explicitly disclaims the exact numbers our thesis leaned on:

> *"Sub-category and use-case breakdowns should be treated as directional at best. Subcategories reflect synthesized notes and anecdotal patterns, rather than precise accounting."*

We cite it as an indicative signal and nothing more. The good news is that the thesis survives on far better evidence, and gets sharper in the process.

The replacement thesis, and it is stronger

Fang, Yuan, Zhang, Donati & Sarvary, "Generative AI and Firm Productivity: Field Experiments in Online Retail" (Zhejiang University / Columbia Business School, 2025–26). **Seven randomised controlled trials** on a major cross-border e-commerce platform, individual experiments ranging from 30,000 to **13.7 million** participants.

WORKFLOW	WHO INITIATES	EFFECT ON SALES
Pre-sale service chatbot	Customer	+16.3% ($p<0.01$), conversion +21.7%
Chargeback defence	Customer/event	+15% success rate
Search query refinement	Customer	+2.93% ($p<0.05$)
Product description generation	Customer reads it	+2.05% ($p<0.05$)
Marketing push messages	Firm	No significant effect (13.7 million consumers)
Google ad titles	Firm	No significant effect (1,244,016 products)

A note on this study. Fang and colleagues ran seven separate experiments at a single cross-border retail platform and report them separately; they were never pooled, and the authors do not use the terms customer-initiated or firm-initiated. That reading of the pattern is ours. The paper is an October 2025 working paper and has not been peer-reviewed.

Roughly three million AI-generated personalised outbound messages against two thousand human-written controls produced **nothing**.

> *AI in sales works when it responds to something a customer did. It does not work when it pushes messages at people who did not ask.*

That is causal, enormously powered, and it explains the front-office/back-office asymmetry better than the asymmetry itself does. Every verdict below is downstream of it.

The ten functions

Verdicts: **Proven** · **Promising** · **Unproven** · **Oversold**

---

1. Prospect research and enrichment — **Unproven**

For: Salesforce's State of Sales 2026 (4,050 people, 22 countries) reports 55% of orgs using AI for prospecting and expected time reductions around 34%. Salesforce sells Agentforce, and these are self-reported expectations, not measured outcomes.

Against, and this is the real evidence: two peer-reviewed agentic benchmarks test exactly this task — gather many facts about many entities, assemble a structured table.

WideSearch (ByteDance Seed, ICLR 2026, 200 curated tasks): best system **4.5–5.1% strict success**. The load-bearing number is item-level F1 of **~73%** — roughly one data point in four is wrong — and row-level F1 of **~52%**, meaning about half of assembled records contain an error. Humans scored near 100% with cross-validation. Named failure modes include **misattributing information to the wrong entity** and **hallucinating when search fails**.

DeepWideSearch (Alibaba, 220 questions): best frontier agent **2.39% success, item-F1 32.9%**.

Those are precisely the errors that survive undetected into a prospect list.

Verdict: narrow deterministic enrichment — verifying one field against a known source — is plausibly solved. Autonomous agentic research and list-building measurably is not, and no independent study shows otherwise.

2. Outreach personalisation across the business — **Oversold**

The best-evidenced finding in the entire report, and it contradicts the entire category pitch.

UK AI Security Institute with Oxford, LSE, Stanford and MIT: **76,977 recipients UK adults**, 19 LLMs, 707 issues, 91,000+ conversations. Effect of personalisation: **+0.43 percentage points** [0.22, 0.64]. **No individual personalisation method exceeded one percentage point.**

What did work was information density — the "information" prompt was 27% more persuasive than the basic one, and each additional fact-checkable claim added +0.30pp. But: **GPT-4o's accuracy fell from 78% to 62% when prompted for density, and 81% of persuasive improvements involved systematic accuracy decreases.** The most persuasive AI outreach is the outreach most likely to be false.

Hackenburg & Margetts (PNAS, 2024): GPT-4 microtargeting from real user data was **not statistically different from non-targeted messaging.**

Hölbling, Maier & Feuerriegel (Scientific Reports, Dec 2025), meta-analysis, 17,422 participants: LLMs versus humans on persuasiveness, **Hedges' g = 0.02, p = .530.** No difference.

Ben-Zion & Lazebnik (2026): randomised crossover field experiment, **16,880 real emails, 121 employees, six companies.** AI rewriting produced **no direct change in open rates, reply rates or response times.**

Dubé & Xu (Quantitative Marketing and Economics, Jan 2026), three randomised trials across about 27,500 customers each, is the strongest study showing AI email works — and its finding is **parity** with human-written, not superiority. The business case is cost substitution. It is also warm retail email to an opted-in list, not B2B cold outreach.

Deliverability: Validity's seed-list benchmark shows global inbox placement falling 84.8% → 83.5%, with spam placement nearly doubling within 2024 from 4.5% to 8.6%,

attributed partly to AI-generated volume. Validity sells deliverability tooling and measures permission-based mail only, so it likely understates the effect for cold outreach. Structurally, Google's mandatory **0.3% spam-rate ceiling** (February 2024) is now the binding constraint on volume, not generation cost.

And the trust penalty: Schilke & Reimann (*OBHDP*, 2025), **13 experiments** — disclosing AI authorship reduces trust, mediated by reduced perceived legitimacy, reliable across framings.

Verdict: no independent measurement of AI cold outreach performance exists in either direction. Every "3x reply rate" claim is vendor-authored. The nearest rigorous analogues put the personalisation premium at roughly half a percentage point.

---

3. Follow-up and nurture — **Oversold**

The single largest test of the core proposition — Fang et al.'s push-messaging arm, on **13.7 million consumers** — returned +1.6%, not significant. The meta-analysis says LLMs are no more persuasive than humans. The one controlled government evaluation (UK Department for Business and Trade, 1,000 staff **with a control group**) found email drafting saved **0.2 hours per task, the lowest of any task measured**, and stated plainly: *"The evaluation did not find evidence that time savings have led to improved productivity."*

Kim & Han (*Behavioral Sciences*, Sept 2025, 360 customers): under high privacy concern, maximum PII-based personalisation performed **no better than a generic message** — a backfire effect in exactly the privacy-sensitive B2B contexts where AI nurture is sold.

Deal recovery rates: no credible independent measurement exists. Every figure in circulation originates from a company selling outbound tooling.

Verdict: reactive follow-up — responding when a buyer does something — sits on the +16.3% side of that split category. Firm-initiated nurture sequences sit in the null category. The distinction is the whole game and the marketing collapses it.

4. CRM hygiene and data capture — **Promising**

The best-evidenced function here, and the one where the operational thesis holds.

No direct study of AI CRM capture exists. But **ambient clinical documentation** is the same technical task — listen to a conversation, generate a structured record into a system of record — and healthcare demanded evidence from randomised trials.

Lukac et al., *NEJM AI* 2025: three-arm pragmatic randomised trial, **238 physicians**, 14 specialties. Product A: **−9.5% time-in-note** (p=0.02). **Product B: −1.7%, p=0.66 — no significant change.** Two products doing the same job; one delivered, one did not. **Returns are vendor-specific, not category-specific.**

Olson et al., *JAMA Network Open* 2025 (263 clinicians, 6 health systems): burnout 51.9% → 38.8% (OR 0.26), after-hours documentation −0.90 hours. No control group.

The error profile matters. Koenecke et al. (ACM FAccT 2024): **~1% of transcriptions contained entirely hallucinated content**, 38% of which carried explicit harms. Asgari et al. (*npj Digital Medicine*, 2025) — 450 transcript-note pairs, 12,999 sentences, **50 doctors annotating** — found a **1.47% hallucination rate** with 44% classed major, and a **3.45% omission rate** where 55% of major omissions concerned *current issues*, the most decision-relevant material. **30% of hallucinations were negations** — stating the opposite of what was said.

Verdict: modest, real, tool-dependent time savings, with a non-trivial error rate concentrated in the worst possible places. Not "transformed data quality."

---

5. Meeting prep and briefing — **Unproven, with a specific evidenced risk**

No study of any kind measures whether AI pre-call briefs improve any sales outcome. Complete vacuum. Every result is vendor marketing.

The closest experimental analogue is alarming. **Dell'Acqua et al., 758 BCG consultants**, three-arm randomised trial. On a task requiring synthesis of quantitative data with qualitative interview insight — structurally identical to a brief built from CRM records plus email history:

CONDITION	CORRECT
No AI	84.5%
GPT-4	70.0%
GPT-4 + prompt training	60.0%

AI users were 19 percentage points less likely to be right — and training made it worse — while simultaneously producing better-presented output (+17.9% to +25.1% on quality ratings).

The counterweight: **Dell'Acqua et al., "The Cybernetic Teammate," NBER 33641, 776 professionals at Procter & Gamble.** Individuals with AI matched two-person teams without it. AI can substitute for absent preparatory expertise.

Verdict: plausible upside, unmeasured, and the specific failure mode is confident-and-wrong output consumed minutes before a meeting with no opportunity to verify it.

---

6. Call recording, transcription and coaching — split verdict

Transcription and admin capture: Proven (time saved), with the caveat below.

Coaching: PROMISING, conditionally. Luo, Qin, Fang & Qu, *Journal of Marketing* 85(2), 2021 — three randomised field experiments, **429 / 100 / 451 agents**. The benefit of an AI coach follows an **inverted U**: middle-ranked agents gain most, bottom-ranked gain little through information overload, top-ranked through algorithm aversion. Throttling feedback volume substantially improved bottom-tier results. **A hybrid AI-plus-human arrangement beat either alone.**

Monitoring as a mechanism: Disproven. Conversation intelligence is, mechanically, electronic performance monitoring plus automated feedback. **Ravid et al., *Personnel Psychology* 2023: meta-analysis, 94 separate studies covering 23,461 people** — monitoring's effect on task performance is **effectively zero** (95% CI −.06 to .06), while stress rises and invasive monitoring raises counterproductive behaviour . **Siegel, König & Lazar (2022), 70 independent samples, independently replicate the null** (performance r = −.01).

Win rate, quota attainment, ramp time: Unproven. No independent study of any commercial platform — Gong, Chorus, Clari, Avoma — on any of these outcomes exists. The most-cited work in the category, "We Analyzed 25,537 Sales Calls," is a vendor analysing its own customer base with no control group.

Cost, from Vendr's anonymised procurement data (593 transactions): median contract **$54,950 per year**, range $11,218–$204,025. A category with nine-figure vendors and $55k median contracts, and no independent outcome evidence.

Emerging cost nobody prices: wiretap and CIPA class actions. *Taylor v. ConverseNow* (N.D. Cal., motion to dismiss **denied 11 August 2025**) held an AI vendor to be a third-party interceptor because it used call data to improve its own service. Statutory damages under the federal Wiretap Act run to $10,000 per violation per class member.

---

7. Lead routing and speed-to-lead — **Unproven**

The most report-worthy finding in the section: a large product category rests on two studies from 2007 and 2011, sharing a lead author and a single vendor sponsor, never peer-reviewed, never independently replicated.

The "5-minute rule" and its 21x/100x multipliers come from Oldroyd & Elkington (2007): **six companies**, drawn from three years of **InsideSales.com's own system data**, funded by InsideSales.com, which sold lead-response software. The 2011 *Harvard Business Review* follow-up (2,241 companies audited; 1.25M leads) shares co-author

David Elkington, then CEO of InsideSales. HBR is not peer-reviewed. The data predates ubiquitous smartphones.

No independent replication exists in fifteen years. Searches structured for academic literature returned only vendor blogs.

That places speed-to-lead at **Unproven** under this report's definitions rather than Oversold: no independent evidence exists either way. Unproven is a statement about the literature. Firms that answer enquiries quickly may well win more of them; nobody outside the vendors has measured it. What fails verification is the specific multiplier, not the practice.

Both studies are observational. Fast-responding firms are plausibly better-resourced and buying better leads; neither can separate response speed from general organisational quality. The 2007 study measured **first call attempt, not contact** — conflating lead reachability with firm responsiveness.

Zombie statistics — do not cite: "78% of customers buy from the company that responds first" (no original source, circular vendor citation); "391x more likely to convert" (unverifiable); "35–50% of sales go to the first responder" (InsideSales-authored).

AI qualification itself: PROMISING. Luo, Tong, Fang & Qu, *Marketing Science* 38(6), 2019 — randomised field experiment, **6,255 calls**. Undisclosed AI chatbots matched **top-quintile human agents** and were ~4x more effective than inexperienced humans (24.8% vs 4.9% purchase rate). But **disclosing AI identity up front cut purchase rate by 79.7%** and drove hang-ups to 56.3%. As disclosure norms and regulation tighten, that advantage may substantially erode. The study is pre-dates current models (2019).

Routing logic specifically: Unproven. There is no research literature of any kind — for or against — on whether ML-based rep-lead matching beats round-robin.

8. Proposal and quote generation — **Unproven** (documents) / **Promising** (pricing)

No independent measurement of AI proposal, SOW or RFP generation exists on time-to-proposal, quality, or win rate. Every figure in circulation is vendor-authored.

The pricing half is different. **Karlinsky-Shichor & Netzer, *Marketing Science* 43(1), 2024** — embedded randomised field experiment at a US metals distributor, 2,075 quotes, alongside observational data on **67,851 quotes / 139,869 pricing decisions**. Profit per line **$105.11 vs $94.16, roughly +11%**, achieved through **higher acceptance at lower prices** — the algorithm corrected systematic over-pricing. Rep compliance was only **19.5%**, so the gain came despite four-fifths of recommendations being ignored. In the counterfactual holdout, **hybrid human-machine (+7.8%) beat full automation (+4.9%)**.

And the finding that should end self-reported time savings as evidence: METR (2025) — randomised trial, 16 experienced developers, 246 real tasks. AI tools made them **19% slower**, while they predicted a 24% speedup beforehand and still believed they had been **20% faster** afterwards. A 39-point perception error, in the only setting where it has been measured against a clock.

Related: "workslop." BetterUp Labs with Stanford Social Media Lab, **1,150 US desk workers** — 40% received AI-generated low-quality work in the prior month, each incident costing an average of **1 hour 56 minutes** of rework, and **half of recipients rated the sender as less capable.** For a document whose entire job is to signal competence to a buyer, that is the relevant risk.

---

9. Sales forecasting — **Oversold**

The clearest gap between marketing and measurement in the report.

What is supported: ML beats the win-probability field a rep types in. **Rezazadeh (*Forecasting*, 2020), 25,578 closed opportunities:** 85% versus 67% accuracy.

Yan et al. (AAAI 2015), 100–200K leads per quarter at a Fortune 500 firm: machine-learning accuracy of 71% to 75% versus salesperson 61% accuracy–0.689. (Four co-authors are IBM Research; IBM sells sales analytics.)

Read the absolute numbers. Best-in-class deal-level accuracy is **0.71–0.75**. That is weak discrimination. The finding is not "AI forecasts accurately" — it is **"AI forecasts poorly, and salespeople forecast worse."**

What is not supported: no independent evaluation of any commercial forecasting or revenue intelligence product exists. Vendor claims of "67% to 98%+ accuracy" carry no methodology, sample, or baseline. **No study links AI forecast accuracy to any business outcome** — all credible evidence stops at accuracy on historical CRM records. And the models are trained on data that only **47% of sales leaders consider high quality**.

The general ML-versus-statistics picture is instructive. In the **M5 competition** (42,840 series, 7,092 participants), the winner beat the best statistical benchmark by 22.4% — but **only 7.5% of teams beat that benchmark at all, and 51.6% failed to beat a naive method.** Improvement was 40% at the most aggregated level and **3% at individual product-store level.** Accuracy gains collapse as you disaggregate — and a B2B pipeline is about as disaggregated, sparse and covariate-poor as forecasting problems get.

---

10. Renewal and churn — **Proven** (prediction) / **Oversold** (retention)

The only function with a deep independent academic literature, and that literature says the standard commercial practice is wrong.

Prediction is real, with a ceiling. KDD Cup 2009 (100,000 customers, 15,000 variables) — the entire global ML community could not push churn accuracy past **~0.77**, and the winning ensemble beat the best single model by 0.014. **Treat any churn accuracy above ~0.85 accuracy as a benchmark artefact:** the 2025 systematic review of 61 peer-reviewed papers (Imani et al.) found reported accuracies

from 84% to 98.8% and **explicitly refused to compare them**, citing incompatible datasets and imbalance ratios. The one peer-reviewed B2B SaaS study reporting 91% accuracy used a dataset that was **54.5% churners** — a constructed balance, not an operational base rate.

Then the causal evidence, which is the important part.

Ascarza, Iyengar & Schleicher (*JMR*, 2016): randomised field experiment at a telecom. Customers flagged at risk were proactively offered a cost-saving plan. **6% of the control group churned; 10% of the treated group did.** The intervention **increased churn by four percentage points.**

Ascarza (*JMR*, 2018), winner of the AMA's Paul E. Green Award: **the customers with the highest predicted churn risk are not the customers who respond to retention offers.** Targeting on responsiveness beats targeting on risk. Most retention programmes fail on targeting.

Verhelst et al. (Orange Belgium, 2023): a real running AI-supported retention campaign, 11,896 customers. Churn **3.6% control versus 3.4% treated.** Two-tenths of a percentage point.

What does work: Lemmens & Gupta (*Marketing Science*, 2020) — ranking by profit rather than churn probability delivered **≥4% profit increase from the same budget**, purely by changing who was targeted. **Retana, Forman & Wu (*M&SOM*, 2016)** — proactive onboarding education, applied universally rather than to predicted-risk customers, **halved first-week churn** and raised eight-month usage 46.6%.

The line for the report: predicting churn and preventing churn are different problems, and the published causal evidence says solving the first one well can make the second one worse.

Summary table

FUNCTION	VERDICT	STRENGTH OF INDEPENDENT EVIDENCE
Prospect research and enrichment	**Unproven**	Strong evidence of failure modes; none of efficacy
Outreach personalisation	**Oversold**	Strong (76,977 recipients) — premium is ~0.43 percentage points
Follow-up and nurture	**Oversold**	Strong (13.7 million consumers, no significant effect)
CRM hygiene and capture	**Promising**	Moderate by proxy (a randomised trial of 238 physicians)
Meeting prep and briefing	**Unproven**	Absent; adjacent evidence negative (–19pp)
Call transcription	**Proven** (time only)	Moderate
Call coaching	**Promising** (conditional)	Moderate (randomised trial, 429 agents/100/451)
Call monitoring as a lever	**Disproven**	Strong (two meta-analyses, 23,461 people + 70 samples)
Speed-to-lead	**Unproven**	No independent evidence either way; the popular multipliers are vendor-funded and unreplicated
AI qualification	**Promising**	Moderate (a randomised trial across 6,255 leads), pre-dates current models, disclosure-fragile
Lead routing logic	**Unproven**	No literature exists
Proposal generation	**Unproven**	None
Quote pricing	**Promising**	Moderate-strong (field experiment)
Forecasting	**Oversold**	Weak; no product ever independently evaluated
Churn prediction	**Proven**	Strong (ceiling about 77% accuracy)
Churn *prevention*	**Oversold**	Strong causal evidence against current practice

Four findings that cut across everything

1. Reactive beats proactive. The cleanest organising principle available, and it is causally identified at one company's seven experiments. Where the customer opened the conversation, AI helped. Where the firm opened it, it did not. That company sells to consumers, so applying the principle to B2B is a transfer the evidence does not itself make — see *Who this evidence does not cover*.

2. The gains land at the bottom of the distribution, consistently. BCG consultants: +43% below-average versus +17% above-average. Support agents (Brynjolfsson, Li & Raymond, *QJE* 2025, **5,179 agents**): +14% average, **+34% novices**, minimal for the experienced. Metals pricing: $14.85 per quote for low-expertise reps versus $5.21 for high-expertise. AI coaching: inverted U, mid-tier gains most. **AI redistributes performance rather than lifting everyone. The reliable business case is levelling up the bottom of a team.**

3. Hours saved is not a business result, and every rigorous study that looked for the conversion failed to find it. The UK DBT evaluation found no productivity gain, with control-group corroboration. Gartner found **72% of sales organisations fail to reinvest** saved time. And the savings estimate itself falls as rigour rises: cross-government self-report 26 min/day, DWP 19 min/day, **HMRC's randomised cohort design with a bias correction: 12 min/day.** A factor of two.

4. The most important single number, and it is from an unimpeachable source. Gartner, 210 chief sales officers and senior sales leaders, early 2026: **25% of sales organisations report ≥50% positive ROI on AI. 20% report ≥50% negative return.** One in five is measurably worse off.

Sourcing notes

Excluded by name: Forrester Total Economic Impact studies. They are vendor-commissioned, use vendor-nominated interviewees, and model a synthetic "composite organisation." Every TEI study encountered (Clari, Writer, boost.ai, Microsoft) was explicitly commissioned by the vendor being assessed. They circulate widely as though independent.

Analyst-brand laundering. Gartner's own CSO Conference release of 20 May 2026 carries at least four outcome statistics — including "55% reduction in onboarding ramp time" and "91% more meetings booked within two months of deploying AI SDRs" — with **no named study, sample size or methodology.** These are already circulating in vendor marketing as "Gartner found." Naming this mechanism is part of the report's job.

Vendor-authorship rate. Across roughly sixty targeted searches, **essentially 100% of quantified commercial-outcome claims** for these functions were vendor-authored or vendor-derived.

Part III — Why most of it fails

Seven failure modes recur across the evidence. None of them is a technology problem.

1. Automating the visible task instead of the expensive one

The tasks that irritate people are visible, repetitive and present. The failures that cost money are omissions — the follow-up that never went out, the signal nobody read, the account nobody chased. Omissions have no friction, so nothing draws attention to them. Organisations systematically automate the first category and leave the second.

2. Pushing when they should be responding

The single best-identified finding in the field. **Fang et al.'s seven randomised trials** show a pre-sale chatbot producing +16.3% on sales among 44,614 consumers, while push messaging to 13.7 million produced +1.6%, not significant. Most sales AI budget is spent on the second category.

3. No baseline, so nothing can be proven either way

Gartner (227 chief sales officers, August–September 2025): 31% of chief sales officers name difficulty proving the ROI of AI-driven tools as a top challenge for 2026. Deployments begin without a measured pre-state, which means

neither success nor failure can be established afterwards, and the decision to continue becomes a matter of belief.

4. Time saved is never converted

Three independent sources converge. The **UK Department for Business and Trade evaluation** (1,000 staff, with a control group) states plainly that *"the evaluation did not find evidence that time savings have led to improved productivity"*, and that control-group colleagues observed no improvement in participants. **Gartner: 72% of sales organisations report low reinvestment of AI-saved time into high-value activity.** MIT NANDA found gains came "without material workforce reduction… tools accelerated work, but did not change team structures or budgets."

Hours saved is an input. Almost nobody builds the management process that turns it into an output.

5. Self-reported time savings are not real savings

METR (2025): a randomised trial of 16 experienced developers across 246 real tasks found that AI tools made them **19% slower**, while they had predicted a 24% speedup beforehand and still believed afterwards that they had been **20% faster.** A 39-point perception error, in the only setting where the question has been measured against a clock rather than a survey.

Corroborating this, the size of reported savings falls as methodological rigour rises: UK cross-government self-report 26 minutes a day; DWP 19 minutes; **HMRC, using a randomised cohort design and applying an explicit ~20% correction for non-usage and respondent bias, 12 minutes.**

6. Monitoring mistaken for enabling

Ravid et al., *Personnel Psychology* 2023: meta-analysis, 94 separate studies covering 23,461 people. Electronic performance monitoring has **no effect on task performance at all**, raises stress, and where invasive raises counterproductive work behaviour. **Siegel, König & Lazar (2022), 70 independent samples**, replicate the

null. Observation is not improvement, and a great deal of sales AI is observation sold as improvement.

7. Buying on evidence that does not exist

Gartner, June 2025: of the thousands of vendors describing themselves as agentic AI, roughly 130 are genuine — the rest is "agent washing" of existing chatbots, assistants and robotic process automation. Gartner forecasts **over 40% of agentic AI projects will be cancelled by end-2027**, citing escalating costs, unclear business value and inadequate risk controls. Its **2025 Hype Cycle for Revenue and Sales Technology places "AI Agents for Sales" at the Peak of Inflated Expectations**, noting vendor investment has created expectations that outpace current capability.

Buyers cannot price-compare or evaluate a category in which most of the labelling is inaccurate and none of the outcome claims are independently verified.

Part IV — What the minority who succeed do differently

Six patterns, each supported by controlled evidence.

1. They respond rather than push

The organising principle of the whole report. Deploy where a customer has already done something — submitted a form, asked a question, arrived on a page, booked a call. **+16.3% versus a non-significant +1.6%**, in the same programme of experiments at the same company — a consumer retail platform, which is a different buying population from the B2B one this report addresses.

2. They aim at the bottom of the skill distribution

The most reliably replicated finding in this literature, across four independent studies in different domains:

STUDY	POPULATION	FINDING
Brynjolfsson, Li & Raymond, *QJE* 2025	5,179 support agents	+14% average, **+34% novices**, minimal for experienced
Dell'Acqua et al., BCG	758 consultants	**+43% below-average**, +17% above-average
Karlinsky-Shichor & Netzer, *Marketing Science* 2024	17 reps, 67,851 quotes	**$14.85/quote low-expertise** vs $5.21 high-expertise
Luo et al., *Journal of Marketing* 2021	980 agents across 3 experiments	Inverted U — mid-tier gains most

AI redistributes performance rather than lifting everyone. The dependable business case is levelling up the bottom of a team. Organisations that deploy it as a uniform productivity multiplier are measuring against the wrong expectation.

3. They build hybrids

Pricing: in the counterfactual holdout, full automation delivered +4.9% profit versus humans; **human-machine hybrid delivered +7.8%, and beat full automation by 3.1 points.** Humans held the advantage on extreme-cost quotes, multi-line quotes, and anywhere private information mattered.

Coaching: Luo et al. found a **hybrid AI-plus-human arrangement outperformed either alone**, and that restricting the AI's feedback volume substantially improved outcomes for the weakest agents, who otherwise gained least, because they hit information overload.

More autonomy is not more value. In both controlled tests, the mixed design won.

4. They target on who responds

The costliest and least intuitive finding in the report. **Ascarza (*Journal of Marketing Research*, 2018)**, AMA Paul E. Green Award winner: **the customers with the highest predicted churn risk are not the customers who respond to retention offers.** Most retention programmes fail on targeting, not on incentive size.

Ascarza, Iyengar & Schleicher (*JMR*, 2016): a proactive intervention aimed at at-risk customers **raised churn from 6% to 10%** in a randomised field experiment.

Lemmens & Gupta (*Marketing Science*, 2020): ranking customers by profit-weighted return rather than churn probability produced a **≥4% increase in overall firm profit from a single campaign, at identical spend, purely by changing who was targeted.**

5. They intervene universally where the intervention is good

Retana, Forman & Wu (*M&SOM*, 2016): at a major cloud infrastructure provider, 2,673 new customers with 366 treated, a simple proactive onboarding-education intervention **halved first-week churn**, cut week-one support questions by 19.6%, and raised *accumulated* usage over eight months by 46.6%.

Note what did the work: a well-designed intervention applied to everyone. No model predicted who needed it.

6. They keep humans on anything that requires judgment

Dell'Acqua et al.: on tasks requiring synthesis of quantitative data with qualitative context, AI users were **19 percentage points less likely to be correct — while producing better-presented output.** Confident, well-formatted and wrong is the characteristic failure, and it is invisible at the point of consumption.

Schilke & Reimann (*OBHDP*, 2025), 13 experiments: disclosing AI authorship reduces trust, mediated by reduced perceived legitimacy. **Luo et al. (*Marketing Science*, 2019):** disclosing AI identity before a sales conversation cut purchase rate by **79.7%** and drove hang-ups to 56.3%.

Draft-first can do two jobs at once, and organisations that treat it only as a safety measure get half the value. It bounds the risk on customer-facing work, and it is the point at which the operator's judgement becomes available as training signal: every edit made before sending is a correction the model could learn from.

That second job requires building. Edits do not propagate on their own — they have to be captured, converted into examples, instructions or evaluation cases, and fed back. Where that loop exists, what sits inside the boundary can widen as calibration accumulates. Where it does not, draft-first is a review queue. No independent study has measured the effect of such a loop either way.

Where the boundary settles is a design decision, and the evidence above says it should not settle at full autonomy for anything a customer sees.

Part V — What it actually costs

There is no independent cost benchmark for AI in any sales function. Not one. Every "average cost" figure in circulation comes from a vendor, an analyst note without a stated method, or a survey of budget intentions rather than spend. What follows is the closest available substitute: actual contract data, disclosed list prices, and the categories of cost that buyers reliably fail to count.

The line item is the smallest number in the equation

Vendr publishes ranges from executed customer contracts rather than list prices. For the best-known category in this report:

PRODUCT	MEDIAN ANNUAL CONTRACT	OBSERVED RANGE
Gong (conversation intelligence)	$54,950	$11,218 – $204,025
LeanData (lead routing)	$22,908	—
Chili Piper (scheduling / routing)	$13,500	—

An **18x spread** on the same product is not a pricing table. It is a negotiation, and the position you negotiate from is whether you can describe the outcome you are buying. Nothing in Part II suggests most buyers can.

For general-purpose assistants the pricing is public and flat: **Microsoft 365 Copilot at $30 per user per month**, seat-based, which for a 40-person commercial org is $14,400 a year before anyone has established that the seats get used. Recall the Wharton/GBK finding: **Marketing and Sales has the lowest daily generative-AI usage of any function, at 34%.** Seat-based pricing bills the other 66%.

The software is roughly half the cost, at best

Wharton Human-AI Research with GBK Collective (~800 US enterprise decision-makers, firms with 1,000+ employees and $50M+ revenue, June–July 2025): **67% of these organisations run generative-AI budgets of $5M or more**, and approximately **44% of that budget goes to people and change management rather than to software.**

That ratio is the most useful cost number in this report. It means a licence quote is not a budget; it is an input to a budget roughly twice its size. It also explains the Part III failure modes directly — the 44% is exactly the spend that converts time saved into output, and it is the first line cut when a deployment is justified on licence cost alone.

For scale at the top end, **KPMG Global AI Pulse Q1 2026** (2,110 C-suite leaders, 20 countries) reports an **average enterprise AI budget of $186 million**, against **8% reporting tangible ROI**.

The cost nobody prices: recorded conversations

Conversation intelligence is the one function in Part II with a **Proven** verdict, and it carries a legal exposure that does not appear on any vendor's pricing page.

In **Taylor v. ConverseNow Technologies**, a federal court **denied the defendant's motion to dismiss on 11 August 2025**, allowing wiretap claims to proceed over AI processing of customer conversations conducted without adequate consent. Under the California Invasion of Privacy Act, statutory damages run at **$10,000 per violation, per class member.** A conversation-intelligence deployment recording a few thousand calls in a two-party-consent state is not carrying a $55,000-a-year risk.

This is a consent and disclosure problem, and it is solvable — but it is solvable in advance, by counsel, and not after the class is certified. Every organisation in this report's audience running call recording should be able to say in one sentence how consent is captured and where it is logged.

What we could not find

We looked specifically for: cost per qualified lead under AI-assisted prospecting; total cost of ownership for an AI SDR deployment including the human review time it generates; and any before-and-after cost comparison for a sales function published by anyone other than the vendor supplying the tool. **None of these exist in the independent literature.** Buyers are pricing this category on vendor arithmetic alone.

Part VI — What changes in 2027

This section is opinion, and is marked as such. Everything before it is evidence. What follows is our reading of where the evidence points, and it should be argued with.

1. The evidence gap closes from the bottom. The functions that will get proven in 2027 are the ones already measurable inside the buyer's own system — inbound response time, quote outcomes, onboarding activation, data completeness. They do not require a vendor's cooperation to evaluate, which is precisely why they will get evaluated first. Outreach reply rates and forecast accuracy will stay unproven for the same reason they are unproven now: the only parties who can measure them are the parties who profit from the answer.

2. Agent washing gets priced in. Gartner's forecast that **over 40% of agentic AI projects will be cancelled by end-2027** is, on the evidence in Part III, conservative rather than alarmist. The correction will not be a collapse in spend; it will be a shift in what buyers demand at the point of purchase. The question "what did it do to the number, measured how, against what baseline" is currently rare. By 2027 it will be standard, and roughly 130 genuine vendors will be able to answer it.

3. Seat-based pricing comes under pressure. A pricing model that bills for access while independent measurement shows 34% daily usage is unstable once buyers start measuring. Expect consumption and outcome-linked pricing to be offered — first as a concession in negotiation, then as a differentiator.

4. The hybrid design wins the argument. Both controlled tests in Part IV that compared automation against a human-machine hybrid were won by the hybrid, by meaningful margins, and the disclosure research gives customer-facing autonomy a hard commercial ceiling. The 2027 version of a good deployment is not more autonomous. It is better bounded — the model does the preparation, the human holds the judgment and the relationship, and the boundary between them is designed rather than inherited.

5. Levelling up beats scaling out. Four independent studies say AI redistributes performance toward the bottom of the skill distribution. The organisations that get returns in 2027 will be the ones that stopped buying it as a headcount substitute and started deploying it as a floor-raiser — which is a management change, and is why most will not.

6. Consent becomes a purchasing question. After ConverseNow, recorded-conversation tooling acquires a compliance diligence step it did not have in 2025. This is a good thing and it will slow nobody down who was doing it properly.

Appendix — Method and sources

Method

Four independent research passes were run for this report, each instructed to use primary sources only, to exclude anything authored or commissioned by a company selling the capability being assessed, to name the study, sample size and date for every figure, and to report evidence vacuums as findings rather than filling them with the best available vendor claim.

Vendor-commissioned economic-impact studies were excluded categorically, including by name any Forrester Total Economic Impact study, all of which are commissioned and paid for by the vendor assessed.

Where a widely cited figure could not survive verification, it is corrected in place rather than dropped quietly. The MIT NANDA correction opening Part II is the principal example.

Principal independent sources

Surveys and official statistics

- McKinsey, *The State of AI: Global Survey* 2026 — 1,719 respondents, 97 nations, fielded 4 May – 8 June 2026
- KPMG *Global AI Pulse*, Q1 2026 — 2,110 C-suite leaders across 20 countries
- BCG *2026 AI Radar*
- IBM CEO Study 2026
- Gartner CSO survey, 210 chief sales officers, fielded January–February 2026
- Gartner CSO survey, 227 chief sales officers, fielded August–September 2025
- US Census Bureau, Business Trends and Outlook Survey / CES Working Paper 26-25 — nationally representative, reference period November 2025 – January 2026
- Wharton Human-AI Research with GBK Collective — ~800 US enterprise decision-makers, June–July 2025
- UK Department for Business and Trade evaluation — 1,000 staff, control group
- HMRC randomised cohort evaluation; DWP evaluation

Randomised and quasi-experimental field evidence

- Fang et al. — seven separate randomised field experiments at one cross-border retail platform, arms ranging from ~30,000 to 13.7 million consumers. Working paper, October 2025, not peer-reviewed
- Brynjolfsson, Li & Raymond, *Quarterly Journal of Economics* 2025 — 5,179 support agents
- Dell'Acqua et al. (BCG / Harvard) — 758 consultants
- Luo, Tong, Fang & Qu, *Marketing Science* 2019 — AI disclosure field experiment
- Luo et al., *Journal of Marketing* 2021 — 980 agents, three experiments
- Karlinsky-Shichor & Netzer, *Marketing Science* 2024 — 17 reps, 67,851 quotes
- Ascarza, *Journal of Marketing Research* 2018 — AMA Paul E. Green Award
- Ascarza, Iyengar & Schleicher, *JMR* 2016 — randomised retention experiment
- Lemmens & Gupta, *Marketing Science* 2020
- Retana, Forman & Wu, *M&SOM* 2016 — 2,673 customers, 366 treated
- METR 2025 — 16 experienced developers, 246 tasks

Meta-analyses

- Ravid, Tomczak, White & Behrend, *Personnel Psychology* 2023 — 94 studies covering 23,461 people
- Siegel, König & Lazar 2022 — 70 independent samples

- Schilke & Reimann, *OBHDP* 2025 — 13 experiments

Market and cost data

- Vendr executed-contract data
- Microsoft published list pricing
- Gartner, June 2025 — agentic vendor census and cancellation forecast
- Gartner *2025 Hype Cycle for Revenue and Sales Technology*

Legal

- *Taylor v. ConverseNow Technologies* — motion to dismiss denied 11 August 2025

Cited with stated caveats, not relied upon

- MIT NANDA, *The GenAI Divide: State of AI in Business 2025*, July 2025

The eleven evidence vacuums

Stated as findings. Each was searched for specifically and does not exist in independent literature.

1. Reply or meeting rates from AI-personalised cold outreach in B2B
2. Cost per qualified lead for AI-assisted prospecting
3. Deal recovery rates from AI follow-up
4. CRM data-completeness improvement from automated capture
5. Any outcome measurement whatsoever for AI meeting preparation
6. Any independent evaluation of a commercial conversation-intelligence platform
7. Any replication of the speed-to-lead research since 2011
8. Any research literature at all on AI lead-routing logic
9. Any independent evaluation of a commercial AI forecasting product, or any study linking AI forecast accuracy to a business outcome
10. Any independent cost benchmark for any function in this report
11. Any B2B replication of the customer-initiated versus firm-initiated result — the report's organising principle rests on a consumer retail platform
12. **Any measurement of AI's effect in an organisation that did not previously staff the function at all** — the entire literature measures substitution against an existing team