PARTNER IN GROWTH
LONDON · UK & EUROPE

THE PLAYBOOK · EDITION 02

The AI Marketing Playbook for *B2B SaaS.*

A practical guide to building AI-assisted marketing workflows without replacing your team — the operating principles, the workflow library, and the governance that keeps your brand out of the failure statistics.

7 WORKFLOWS | 4 PRINCIPLES | 24 DIAGNOSTICS | 90-DAY ROLLOUT

Hilal Tasdan
FRACTIONAL CMO · PARTNER IN GROWTH

PARTNERINGROWTH.CO.UK

Nearly every B2B marketing team now uses AI. Fewer than four in ten can show it working.

Ninety-five per cent of B2B marketers use AI tools. Eighty-seven per cent report productivity gains. And only **39% can show any improvement in how their content actually performs** (Content Marketing Institute, n=1,015, 2026). That gap — between feeling faster and being better — is what this playbook is about.

The teams on the right side of the gap did not buy better tools. They built three things the others did not: **a context layer** the AI works from, **workflows with deliberate human checkpoints**, and **governance** that makes quality enforceable rather than aspirational. Tools are the cheapest and least important part of the stack. This document covers the other three.

WHO THIS IS FOR

- B2B SaaS marketing teams of one to fifteen people
- Founders being told to "do more with AI" without a plan for what that means
- Teams whose AI usage grew bottom-up and now needs structure
- Anyone deciding what their team should stop, keep and start

WHAT THIS IS NOT

- A tool list — tools change monthly; the workflows and principles here do not
- A case for cutting your team — the evidence points the other way
- A prompt pack — prompts encode individual skill; systems encode organisational capability
- Hype — every claim here carries its source, and the shaky ones are flagged

On "without replacing your team." That phrase in the subtitle is a position, not a platitude. The best evidence available — a controlled study of 758 professionals — found AI lifted output quality most for the *least* experienced people, and that accuracy fell when humans stopped checking the machine. AI narrows skill gaps and widens judgement gaps. The teams that win keep the judgement in-house.

Contents

TIME TO READ

90 minutes

Or one workflow at a time, as you build each.

TIME TO IMPLEMENT

One quarter

The 90-day plan on page 28 sequences it.

PREREQUISITE

None

Except honesty about what your team actually does with AI today.

The operating principles.

Four rules explain most of the gap between teams that feel faster and teams that perform better. None of them is about which tool to buy, and all of them survive the next model release.

The evidence, honestly read

AI marketing is drowning in numbers, most of them recycled from vendor surveys. Here is what the credible data actually shows.

95%
of B2B marketers now use AI tools — adoption is effectively saturated (CMI, n=1,015, 2026)

87%
report productivity gains — the feeling of speed is nearly universal (CMI, 2026)

39%
can show improved content *performance* — the number that matters (CMI, 2026)

FINDING	NUMBER	SOURCE
AI budget share of marketing budgets	15.3% average; 21.3% at AI-ready orgs	Gartner CMO Spend Survey, n=401 (2026)
CMOs calling AI critical vs ready to scale it	70% vs 30%	Gartner (2026)
AI integrated into daily workflows	19% – 54% use it ad hoc	CMI (2025)
Quota attainment, high vs low AI-adopting GTM teams	67% vs 59%	ICONIQ State of GTM, n≈150 (2026)
GTM headcount at the same ARR, high AI adopters	21–43% smaller orgs	ICONIQ (2026) — correlation, not proven cause
Marketers publishing AI output without review	4%	HubSpot State of AI (2025) — vendor-fielded
Marketers with high trust in AI outputs	4%	CMI (2025)
Teams saying content quality *decreased* with AI	12%	CMI (2026)

THE HONEST SUMMARY

Everyone has the tools. Almost nobody has the operating system. The ICONIQ data is the strongest B2B-specific signal — high-AI-adopting GTM teams post 8 points better quota attainment and run materially leaner — but note the caveat: companies good at AI adoption tend to be well-run generally, and their cost per lead is actually *higher*. AI amplifies an operating system; it does not substitute for one.

THREE NUMBERS TO STOP REPEATING

"95% of AI pilots fail." The MIT NANDA report measured P&L attribution of custom enterprise pilots within months — the same report found 90% of employees use AI personally with strong results.

"10× productivity." No primary source exists. The best controlled study found +12% output, +25% speed, +40% quality.

"Search volume will drop 25%." A 2024 Gartner *prediction*, widely cited as a measurement. It did not materialise on schedule.

Principle 1
Build a system, not a *stack.*

Tool-first adoption produces the martech world's Frankenstack — a fragile chain of subscriptions nobody can debug, each encoding one person's prompting habits. System-first adoption compounds.

A STACK

- Each person prompts their own way, in their own tools
- Quality depends on who did the work that day
- When positioning changes, nothing downstream updates
- When your best prompter leaves, the capability leaves
- Every new piece of work starts from a blank page

This is the 54% "ad hoc" usage in the CMI data — and it caps out at time savings, which is why productivity gains outrun performance gains 2:1.

A SYSTEM

- Shared, versioned context every workflow reads (page 08)
- Defined workflows with named human checkpoints
- Quality rules encoded and enforced automatically
- Outputs measured, and the measurements fed back in
- Every fix improves every future output

"Don't stop at workflows — build a compounding content system" is the Animalz formulation, and it generalises to every function in Part II.

THE TEST

If a new hire ran your best prompt tomorrow, would they get the same output quality?

If not, your prompts encode individual skill, not organisational capability. That is the single cleanest test of stack versus system — and the reason this playbook spends one page on tools and twenty on everything around them. A related test: when your positioning last changed, how many prompts and workflows had to be manually hunted down and updated? In a system, the answer is one file.

> **Where the money goes wrong.** B2B teams are currently five times more likely to increase AI tool spend than to invest in training the humans who operate it — 45% versus 9% (CMI, 2026). Given that the biggest measured performance differences come from how teams work rather than what they buy, that ratio is close to exactly backwards.

Principles 2, 3 and 4

PRINCIPLE 2

Delegate by verifiability × blast radius — not by difficulty

Ethan Mollick's "jagged frontier": AI capability is uneven and unintuitive, so what to delegate must be discovered per task, not assumed. The working rule that emerges from documented teams is a two-axis test: how easily can a human verify the output, and how much damage does an error do before it is caught?

DELEGATE FULLY, SAMPLE-CHECK
Easily verified, low blast radius: enrichment, transcription, dedup, internal summaries, first-pass research.

AI DRAFTS, HUMAN EDITS
Medium: blog drafts, ad variants, nurture copy, repurposing. The default mode for most marketing work.

HUMAN LEADS, AI ASSISTS
Hard to verify or high blast radius: published statistics, pricing and competitor claims, thought leadership, Tier-1 account outreach.

PRINCIPLE 3

Checkpoints are structural, not optional

The BCG/Harvard study of 758 professionals is the most important experiment in this field: with AI, output rose 12%, speed 25%, quality 40% — but on a task designed to fool the model, **human accuracy fell from 84% to 60–70%** because people stopped checking. Reviewers fall asleep at the wheel; that is human nature, not a training gap. The fix is architectural: enforced approval gates at each workflow phase — what the Animalz team calls speed bumps — and AI critics (style checkers, fact checkers) that run before any human sees the draft, so human attention is spent where it is scarce.

PRINCIPLE 4

Context beats prompts

The visible difference between mediocre and excellent AI output is rarely the prompt. It is what the model knows about your ICP, positioning, voice and proof points when it starts. Teams getting real results maintain a persistent, versioned context layer — and every workflow reads from it. Without one, every piece of work starts from scratch and sounds like everyone else's. The next page is the build spec.

CENTAUR OR CYBORG — BOTH, DELIBERATELY

Mollick's two working modes: **centaurs** divide the work cleanly — human does strategy, AI does the research sweep — while **cyborgs** weave back and forth inside a single task, drafting a paragraph, asking for alternatives, editing, asking for a critique. Neither is superior; they suit different work. What matters is that the choice is conscious. Teams that drift into cyborg mode on high-stakes work without noticing are the ones who publish the hallucinated statistic.

The context layer

One shared, versioned home for what your AI needs to know. This is the highest-leverage build in this entire playbook, and it costs a week.

WHAT GOES IN IT

icp.md — segments, triggers, disqualifiers	`Layer 0 of Edition 01`
positioning.md — alternatives, unique value, category	`Dunford components`
messaging.md — hierarchy, claims + their proof	`one claim, one source`
voice.md — rules, each with an exemplar passage	`extracted, not aspirational`
competitors/ — one file each, dated	`refresh quarterly`
proof.md — case studies, numbers, quotes	`the only stats AI may cite`
do-not.md — banned claims, words, comparisons	`legal's page`

Plain markdown files in a shared repo or workspace. Not a wiki nobody reads — the difference is that every AI workflow ingests these files on every run.

THE VOICE FILE — EXTRACT, DON'T ASPIRE

The Animalz method: derive the voice guide from your *actual published corpus* — your ten best pieces — rather than from the aspirational brand document. Style guides describe the voice you wish you had; the corpus contains the one your audience knows. Have AI extract the patterns, a human edit them, and pair every rule with a real exemplar passage. Rules tell the model what to do; examples show it.

TREAT PROMPTS LIKE MODELS

The best documented ABM teams back-test their qualification prompts against accounts with known outcomes before trusting them — the same discipline you would apply to a scoring model. If your ICP-qualification prompt cannot correctly classify last year's twenty best and twenty worst customers, it is not ready to run on ten thousand unknowns.

WHY THIS IS PRINCIPLE 4 AND NOT A NICE-TO-HAVE

Generic output is not a model failure; it is a context failure. The model has seen ten million SaaS blogs and zero of your win/loss calls. Until you fix that asymmetry, more prompting produces faster mediocrity.

Sequencing note. If your ICP and positioning are not written down, the context layer forces the issue — which is why teams that ran the Growth Audit Framework (Edition 01) first find this build takes days, not weeks. The context layer is Layer 0 of that framework, made machine-readable.

II

The workflow library.

Seven workflows, ordered by strength of evidence. Each shows the same anatomy: what the AI does, where the human sits, what the quality gate checks, and what the documented results look like. Adapt the shape, not the specifics.

Anatomy of a working AI *workflow.*

Every workflow that survives contact with reality has the same four stages and the same three-layer review. The failures skip one.

1

Input

The context layer, plus the job-specific brief: keyword, account list, transcript, campaign goal. Garbage here is unrecoverable later.

Human owns: strategy, the brief, what "good" means.

2

Generation

AI produces the draft, the research sweep, the variants, the enrichment — the volume work humans are too expensive for.

AI owns: breadth, speed, first drafts, permutations.

3

Review

AI critics first — style check, fact check, rubric score — then a human gate with a named owner and a checklist.

The load-bearing stage — under-resourcing it is behind most of the failure patterns on page 23.

4

Publish & feed back

Ship, measure, and route the results back into the context layer so the system compounds instead of repeating itself.

Skipping feedback turns a system back into a stack.

THE THREE-LAYER REVIEW, IN ORDER

Layer 1 — AI critiques AI. A second pass — style checker against voice.md, fact checker against proof.md, rubric scorer — runs on every output before a person sees it.	`Cheap`
Layer 2 — Human gate. A named person, a written checklist, and the authority to reject. "The writer should check" is not a protocol.	`Scarce`
Layer 3 — Sampling. For fully-delegated work (enrichment, transcription), audit a random 5–10% weekly rather than reviewing everything.	`Ongoing`

Why this order matters: the BCG finding is that human reviewers degrade when reviewing feels redundant. AI critics do the redundant part — typos, style drift, claim-vs-source mismatches — so the human gate is spent on the things only humans catch: strategy, taste, and whether the piece says anything at all.

Documented extreme: one content agency runs **eight parallel AI agents** reviewing every article line-by-line against an extracted style guide — built in under half an hour. You do not need eight. You need at least one.

W1 Content & SEO production

The best-documented workflow, and the one where the penalty for doing it lazily is now measurable.

1 Brief from the context layer, not from a keyword

The brief inherits icp.md, positioning.md and the target reader's job-to-be-done. A keyword plus "write 1,500 words" is how generic content gets made at scale — and 84% of marketers already admit their campaigns feel generic.

2 AI drafts the structure and the routine sections

Intros, summaries, definitional sections, tables. The human writes the parts that carry the argument: the opinion, the example from a real customer, the section that says something true and slightly uncomfortable. HubSpot's editorial team runs exactly this split and reserves thought leadership for humans entirely.

3 AI critics run before any human reads

Style pass against voice.md. Fact pass against proof.md — every statistic must trace to a source in the context layer or be flagged for verification. Padding pass: cut every sentence that would survive in a competitor's post unchanged.

4 Human editorial gate, with rejection authority

A named editor with a checklist: is the argument true, is it ours, would the ICP forward this? Editing depth is a health metric — if edit time trends to zero, either quality rose or the editor fell asleep; check which before celebrating.

5 Publish, measure, feed back

Performance data returns to the system: which angles ranked, which pages converted, which got cited by LLMs. Winning patterns get added to the brief templates; losing ones get retired.

EVIDENCE FOR THE SPLIT APPROACH

Human-led content holds roughly 80% of Google #1 positions versus ~10% for pure-AI content (Semrush, 42K posts, 2025–26 — an SEO vendor, but data-based) — and 64% of SEO teams now run exactly this human-led, AI-assisted model.

THE PENALTY IS REAL

Google's March 2024 update deindexed scaled AI content sites — the recovery record since is thin, per detection-vendor tracking (read with that incentive in mind). The policy targets *scaled low-value* content, not AI use — intent, not method.

WHERE SCALE DOES WORK

One documented team shipped 1,600+ data-grounded competitor-comparison pages in two months — triple-digit organic growth, #1 rankings on pricing queries. Templated, bottom-funnel, factual: that is the scalable end of content.

W1 continued — the two questions that decide it

QUESTION ONE

What is your unscalable ingredient?

AI makes production free, which makes *what you feed it* the entire differentiator. The teams winning with AI content have a proprietary input: customer call transcripts, win/loss interviews, product usage data, a founder with actual opinions. If your input is the same SERP everyone else is summarising, your output is a commodity at any quality level.

Practical version: before approving any brief, name the ingredient in it that a competitor could not paste into their own prompt. No ingredient, no brief.

QUESTION TWO

Are you writing for readers, buyers, or machines?

The search landscape shifted under this workflow. When an AI Overview appears, clicks to the top result drop by roughly a third (Ahrefs, 300K keywords, 2025); Pew found users click links in 8% of AI-summary sessions versus 15% without. Meanwhile LLM-referred visitors are ~1% of traffic but convert at up to ten times the rate of search visitors (Microsoft Clarity telemetry, 1,200+ sites, 2025 — vendor data).

The implication: informational tofu content is losing its click economics, while being *citable* — original data, clear definitions, named expertise — earns disproportionate value. Publish less, and make what you publish quotable.

QUALITY GATE — W1 CHECKLIST

- ☐ Every statistic traces to proof.md or a named primary source
- ☐ At least one proprietary ingredient a competitor cannot copy
- ☐ Voice pass run against extracted exemplars, not the brand deck
- ☐ A named human editor approved it, with authority to kill it
- ☐ The piece takes a position someone could disagree with
- ☐ Post-publish performance is routed back into the brief library

> **Team note.** This workflow does not remove the writer; it changes the job from typing to editing, arguing and sourcing. The measured risk is deskilling: Microsoft and CMU found higher confidence in AI correlates with less critical thinking. Rotate juniors through the human-written sections deliberately — the judgement you need to review AI in five years is built by writing without it now.

W2 Account research & ABM

The workflow with the best documented economics: research depth per account that was never affordable with humans, at pennies per account.

1 Humans define the ICP; AI qualifies at scale

The team writes the qualification rules — including explicit exclusion signals — and AI research agents classify every company in the TAM from public evidence. The documented teams back-test the qualification prompt against known accounts before running it: if it cannot classify last year's best and worst customers correctly, it does not run.

2 Research agents enrich 30+ data points per account

Tech stack, hiring signals, regulatory events, named stakeholders — from public sources and data providers. Documented cost: a few thousand pounds to map an entire TAM, versus an SDR-year of manual research.

3 A second agent verifies the first

The pattern worth stealing from the best-documented implementation: a dedicated fact-checking agent cross-references the enrichment agent's claims against public sources before anything reaches the CRM. AI checking AI, with humans sampling the residue.

4 Signals trigger plays; tier decides the checkpoint

Funding round, champion job change, competitor complaint — each signal maps to a play. Tier-1 accounts: a human reviews every touch. Long tail: auto-send above a confidence threshold. The tiering is the governance.

DOCUMENTED RESULTS

Meeting bookings (auto-BDR with case-study matching)	3×
Opportunities from the same system	2×
Enrichment coverage with verification agent	~100%
Demo response time at one public SaaS co	24h → 2min

Named-operator teardowns via Growth Unhinged (2025–26). Single companies, not benchmarks — treat as existence proofs, not expectations.

QUALITY GATE — W2

- ☐ Qualification prompt back-tested against known outcomes
- ☐ Verification pass before CRM write
- ☐ Tier rules written down; Tier-1 = human review, always
- ☐ Weekly 5% sample audit of enrichment accuracy

The failure mode here is silent: wrong firmographics quietly poisoning routing, scoring and personalisation for a year. Verification is not optional overhead; it is the product.

W3 Outbound personalisation

The workflow with the strongest results *and* the ugliest documented blowback. The difference is volume discipline.

WHAT WORKS, DOCUMENTED

Closed-lost re-engagement: an agent monitors buying signals on closed-lost accounts, pulls the old call transcript, and drafts an email referencing the actual objection that killed the deal — sent by the rep.

Signal-cluster micro-campaigns: 50–100 contacts who share a live trigger, with copy built from that trigger — not a thousand-row blast with a {{first_name}} token.

Case-study matching: the auto-BDR that books 3× more meetings selects which customer story to send per prospect — relevance, not prose flair, is the lever.

Pipeline share documented at one company: ~25% of new pipeline from AI-assisted outbound with 50%+ open rates. Existence proof, not benchmark.

WHAT BLOWS UP, DOCUMENTED

Deliverability. A 100K-email analysis (vendor-published, read directionally) found AI-written sequences spam-flagged at more than twice the human rate, with fewer meetings per send. Filters detect AI fingerprints; recipients detect laziness. Both compound at volume.

The maths trap. If AI makes a touch 10× cheaper and you respond by sending 10× more, you have spent the entire gain on annoying your market — and rented a reputation problem. The winning teams spend the gain on *depth per touch* at the same or lower volume.

QUALITY GATE — W3

- ☐ Volume cap set *before* the tooling, and enforced
- ☐ Every sequence references a verifiable, current trigger
- ☐ Tier-1 accounts: rep reviews every message, no exceptions
- ☐ Spam-flag and reply rates tracked against a human baseline
- ☐ A defined threshold at which the machine gets paused
- ☐ Suppression list honoured across all tools, not per tool

> **The one-sentence policy that prevents most of it.** “We use AI to earn the right to a smaller, better list — never to afford a bigger, worse one.” Put it in the usage policy verbatim. Every outbound failure case in the research reduces to violating that sentence.

W4 Product marketing & competitive intel

The quiet compounding win: turning positioning, win/loss and competitive work from annual projects into standing capabilities.

1 Build skills, not documents

The best-documented architecture treats each recurring PMM task — win/loss analysis, battlecard refresh, competitor teardown, case-study drafting — as a reusable "skill": a defined procedure that reads specific context-layer files and writes structured output to a defined place. Run monthly, not annually.

2 Feed it primary material

Call transcripts, win/loss interviews, support tickets, community threads. AI is exceptional at synthesising across a hundred transcripts — the work PMMs never had time for. It is unreliable at inventing insight from none; the input corpus is the ceiling.

3 Competitive intel on a refresh cycle

One file per competitor in the context layer, dated. An agent flags staleness and re-runs the teardown — pricing-page diffs, release notes, review-site sentiment — on schedule. Sales gets battlecards that are weeks old, not quarters.

4 Humans own the position

AI can summarise what competitors say and what customers feel. Choosing what your company should claim — the positioning decision itself — is judgement, accountability and nerve. Delegating it produces the median of your category's marketing, by construction.

WHY THIS ONE COMPOUNDS FASTEST

Every other workflow reads from the context layer; this one *writes to it*. Better win/loss synthesis sharpens icp.md; sharper icp.md improves every brief, sequence and battlecard downstream. If you can only build one workflow this quarter, build W1 — it carries the revenue argument. If you can build a second, make it this one.

QUALITY GATE — W4

- [ ] Every claim in a battlecard carries a date and a source
- [ ] Win/loss synthesis quotes real calls, reviewable on request
- [ ] Positioning changes require a human decision, recorded
- [ ] Stale competitor files flagged automatically

W5 Repurposing & social

The easiest workflow to start with and the easiest to do badly. One dense asset becomes twelve; the voice file decides whether the twelve are worth reading.

THE STANDARD SHAPE

One webinar / podcast / long-read in	`transcribed`
AI extracts the 6–10 defensible points	`human confirms`
Per-channel transformation	`pillar post, LinkedIn series, thread,`
Voice + fact pass per output	`against voice.md / proof.md`
Human approves the batch	`one gate, not twelve`
Engagement data returns daily	`winning angles inform the next batch`

The feedback line is what separates this from content confetti: the best documented team pipes channel engagement back into the generation workflow every day.

THE TWO FAILURE MODES

Templatisation. Twelve outputs that all sound like the same LinkedIn ghost-writer. The fix is the extracted voice file plus a human with taste on the batch gate — and the discipline to kill outputs that are technically fine and completely forgettable.

Repurposing emptiness. If the source asset made no argument, twelve derivatives make no argument twelve times. Repurposing amplifies quality in both directions; it never creates it.

A NOTE ON VOLUME

Every channel is flooding with AI content simultaneously; roughly half of consumers say they can tell, and engagement drops when they suspect it. The scarce asset on social is now a recognisable human point of view. Use AI for the formats; keep the opinions.

W6 Analytics & experimentation

The least glamorous workflow and possibly the highest-leverage one: AI as tireless analyst, human as decision-maker.

1 AI does the plumbing

Pulling campaign data, assembling the weekly report, annotating anomalies, drafting the Slack summary and the tickets. This is pure delegation territory: easily verified, low blast radius, endlessly boring.

2 AI proposes hypotheses; humans choose bets

The documented extreme: a four-engineer growth team ran 360 of their company's 514 experiments in twelve months — roughly 1.5 per business day — with AI generating hypotheses from the data and drafting the analysis. Humans picked what to test and what to believe. Their ARR went from $1M to $30M that year. A single company, so an existence proof rather than a benchmark — but the metric that mattered generalises: experiment throughput per person.

3 Spot-check every calculation that reaches a decision

AI arithmetic and AI SQL are good and not reliable. The rule from teams doing this well: any number that will change a budget, a headline or a roadmap gets recomputed once by a human before it ships. Numbers in internal exploration flow free; numbers in decisions get checked.

WHAT CHANGES FOR A SMALL TEAM

The realistic version for a three-person marketing team is not 360 experiments. It is: the Monday report writes itself, anomalies get flagged the day they happen instead of at month-end, and every experiment gets a written hypothesis and a written result because writing them no longer costs an afternoon. That alone moves most teams from 2–3 tests a month to 8–10.

QUALITY GATE — W6

- ☐ Decision-bound numbers recomputed by a human
- ☐ Experiments logged with hypothesis + result, no exceptions
- ☐ AI summaries link to the underlying query
- ☐ Anomaly alerts reviewed same-day by a named owner

W7 Lifecycle email

Last in the library because the public evidence is thinnest — which is itself worth knowing before a vendor tells you otherwise.

THE DEFENSIBLE PATTERN

Humans own the journey architecture — the stages, the triggers, what each moment is for. **AI drafts the variants** inside that architecture: copy per segment, per usage signal, per persona. One documented implementation attributes a 10% feature-adoption lift to signal-driven AI personalisation of lifecycle messages.

Gate anything that touches money. Renewal, pricing, billing and cancellation messages stay human-approved, always — the blast radius of an error there is a churned customer and a screenshot on social.

HONEST EVIDENCE NOTE

Lifecycle and paid-ads AI are the two areas where nearly everything published is written by the vendors selling the workflow. That does not make it wrong; it makes it unverified. Pilot with your own holdout — a control segment that gets your existing emails — and believe your own numbers over anyone's case study, including the ones in this document.

Paid ads, briefly: the same logic applies. AI generates variant volume against your messaging hierarchy; the platform tests; humans gate brand safety and claims. Any "10× faster winners" claim you read is vendor marketing until your own account says otherwise.

THE LIBRARY, PRIORITISED FOR A SMALL TEAM

Start this month — cheapest, fastest proof	W5, W6
Build this quarter — the revenue engines	W1, W2
Build next — compounding + reach	W4, W3
Pilot with a holdout when the rest run	W7

Sequencing logic: W5 and W6 prove the operating model with low risk and visible wins. W1 and W2 carry pipeline. W4 makes everything else smarter. W3 goes late deliberately — it is the workflow where an immature system does public damage. W7 waits for clean journey data.

III

Governance.

The unglamorous part that decides everything. Sixty-three per cent of organisations still have no AI governance at all, and nearly half of B2B marketing teams operate without guidelines while 95% of their people use the tools. The gap between usage and rules is where the incidents live.

The AI usage policy

One page per component, six components, written by marketing with legal — not the reverse. A policy nobody can recall verbatim is a policy that does not exist.

01

Approved uses & tools

Which tools, for which task classes, with which data. The whitelist is less important than the task mapping: the same tool can be approved for drafting and prohibited for publishing customer data.

02

Data rules

What may never be pasted into a model: customer PII, unreleased financials, security details, anything under NDA. Name the categories, not just "sensitive data" — people cannot comply with an adjective.

03

Review requirements by risk tier

The delegation matrix from Principle 2, made policy: what auto-ships with sampling, what needs a human edit, what needs named sign-off. Include who has authority when AI output touches pricing, legal or competitor claims.

04

Disclosure position

Decided deliberately, not by default. The emerging B2B norm: disclose AI imagery and video, disclose in bylined journalism-style content, do not label routine AI-assisted copy — and keep an internal audit trail either way. EU AI Act transparency obligations apply as of August 2026; check them against where you sell.

05

IP & copyright

Purely AI-generated output is not copyrightable under the current US Copyright Office position; human-edited and human-arranged material is protectable in its human parts. If an asset matters commercially, a human must materially shape it — policy and self-interest agree for once.

06

Amnesty & disclosure incentives

Roughly 78% of people who use AI at work bring their own unapproved tools (Microsoft/LinkedIn, n=31,000, 2024). A punitive policy does not stop this; it drives it underground, where governance cannot see it. Make the policy an upgrade path: show us your workflow, keep using it, we will harden it together.

The number that justifies the week this takes. Companies where employees use AI covertly are not avoiding AI risk — they are running it unmanaged. Shadow-AI-related breaches cost measurably more than managed ones, and executives are among the heaviest covert users. The policy's first job is visibility, not restriction.

Brand voice protection

Voice is the first casualty of scale and the hardest thing to recover once your audience decides you sound like everyone else. Three moves protect it.

1

Extract the voice you actually have

Take your ten best-performing pieces — the ones customers quoted back to you. Have AI describe the patterns: sentence rhythm, stance, vocabulary, what it never does. A human edits that description into voice.md. This beats the aspirational brand deck every time, because the corpus contains the voice your audience already recognises.

2

Encode it as rules plus exemplars

Rules alone underdetermine style; examples alone drift. voice.md carries both: eight to twelve rules, each paired with a real passage that demonstrates it, plus an anti-pattern list — the phrases and shapes that mark text as machine-made. Em-dash chains, "in today's fast-paced world," the rule of three in every paragraph, hedges before every claim.

3

Enforce it automatically, on everything

A style-checking pass against voice.md runs on every output before human review — the same pattern as the eight-agent line-by-line reviewer documented at one content agency, at whatever scale you can afford. Voice enforcement that depends on a busy human remembering the guide is voice drift on a delay.

THE QUARTERLY DRIFT AUDIT

Put your five most recent pieces next to the five best pre-AI pieces. Read them aloud. If a stranger could not tell they came from the same company, the system is drifting — and because drift is gradual, nobody inside notices without the ritual. Twenty minutes, once a quarter, calendar-blocked.

WHY IT IS COMMERCIALLY URGENT

About half of readers believe they can spot AI copy, and when they merely *suspect* it, engagement drops and the brand reads as impersonal or lazy. In a feed where every competitor ships machine-flavoured text, a recognisable voice is distribution. Gartner expects a fifth of brands to differentiate on the *absence* of AI by 2027 — a prediction, but a telling one.

The fact-checking protocol

"The writer should check" is not a protocol. Named steps, named owners, and a definition of which claims trigger which checks.

CLAIM TYPE	CHECK REQUIRED	OWNER	WHEN
Statistics & data points	Must trace to proof.md or a named primary source; secondary citations resolved to the original before publishing	Writer, verified by editor	Every output
Product claims	Checked against current product truth — not the roadmap, not last year's deck	PMM or product counterpart	Every output
Competitor claims	Dated evidence in the competitor file; anything older than a quarter re-verified before use	PMM	Every battlecard & comparison
Pricing & legal	Named sign-off; AI-drafted, human-owned, no exceptions	Founder / legal	Every touch
Customer references	Permission on file; quote checked against the transcript	Marketing lead	Every use

WHY THE PARANOIA IS CALIBRATED

On tightly grounded summarisation tasks the best models now hallucinate under 2% of the time — but open-ended factual generation is far worse, and marketing prompts are usually open-ended. The public record of what one invented fact costs: a tribunal ordered Air Canada to honour a refund policy its chatbot made up; Sports Illustrated's fake AI authors made global news, and the publisher's CEO was out within weeks. Marketing's equivalents — an invented benchmark in a comparison page, a fabricated customer stat — are cheaper to prevent than to retract.

THE TWO-LIST TRICK

Maintain proof.md as the *only* pool of statistics AI is permitted to cite, and do-not.md as the list of claims it must never make. Then the fact-check reduces to a mechanical diff: does every number in the draft exist in proof.md? Anything outside the pool gets flagged automatically — which is exactly the kind of tireless, boring checking AI itself is good at. Layer 1 of the review does this before a human ever reads.

Permit uncertainty in every prompt. The single most effective anti-hallucination instruction, straight from the model vendors' own guidance: tell the model it may say "I could not verify this." Models fabricate most when the prompt implies an answer must exist. A workflow that rewards flagged uncertainty over confident invention is safer than any detector.

How AI marketing fails

Eight patterns, all documented, ordered roughly by how often they appear. Every one is preventable with something already described in this playbook.

PATTERN	WHAT IT LOOKS LIKE	THE PREVENTION
1 · Pilot purgatory	Tools bought, demos run, nothing in production. 42% of companies scrapped most AI initiatives in 2025, up from 17% (S&P Global)	One workflow, end to end, with a named owner — before any second tool
2 · Generic flooding	Volume up, differentiation gone; 84% of marketers admit to generic campaigns	The unscalable-ingredient rule (W1); publish less, make it citable
3 · Voice drift	Gradual convergence on template-speak nobody inside notices	Extracted voice file + automated style pass + quarterly drift audit
4 · Published hallucination	An invented stat or claim ships under your brand	proof.md as the only permitted stat pool; fact pass before human review
5 · Outbound blowback	Spam-flag rates double, domain reputation burns, replies drop	Volume caps, tiered review, a written pause threshold (W3)
6 · SEO penalty	Scaled thin content deindexed in a core update; recoveries rare	Human-led editorial, data-grounded scale only, E-E-A-T signals
7 · Shadow AI	~78% of AI users bring their own tools; governance sees none of it	Amnesty-first policy; disclosure incentivised, not punished
8 · Deskilling	Confidence in AI rises, critical thinking falls (Microsoft/CMU, 2025); juniors never build judgement	Rotate humans through unassisted work; edit-depth as a health metric

THE META-PATTERN

Every failure above is a missing *gate*, not a bad model. Purgatory is a missing owner. Flooding is a missing brief standard. Drift is a missing style pass. Hallucination is a missing fact pass. Blowback is a missing volume cap. The model versions will keep changing; the gates are what you actually own.

ONE NUMBER TO WATCH ABOVE ALL

Edit depth per AI draft. If it trends down because quality rose, your context layer is working. If it trends down because reviewers stopped reading, you are four weeks from an incident. Same curve, opposite meanings — and only a human who still reads carefully can tell you which one you are on.

IV

Rollout.

From ad hoc to operating system in one quarter: measurement that means something, a 24-question self-assessment to find your starting point, and a 90-day plan with three builds and a deliberate amount of nothing else.

Measurement that means *something.*

Three tiers, in ascending order of difficulty and credibility. Most teams stop at tier one, which is why 87% report gains and 39% can show them.

TIER 1

Time & throughput

Hours per asset, assets per person, experiment velocity, enrichment coverage. Easy to measure, easy to inflate, and real — as far as it goes.

Necessary but never sufficient: time saved on work that does not perform is efficiency at producing nothing.

TIER 2

Quality

First-pass acceptance rate through the human gate, edit depth per draft, voice-drift audit results, spam-flag rate vs human baseline, error rate in sample audits.

The tier that predicts incidents before they happen. Nobody's dashboard has it; yours should.

TIER 3

Pipeline & revenue

Before/after on the workflow's actual output: organic pipeline from W1, meetings and opportunities from W2/W3, adoption lift from W7 — each against a baseline or holdout.

Three quarters of GTM teams measure AI impact as productivity; only half attempt revenue uplift (ICONIQ, 2026). Tier 3 is where budget arguments are won.

THE DASHBOARD — EIGHT NUMBERS, MONTHLY

- Assets per person per month (Tier 1) ___
- Experiment velocity (Tier 1) ___
- First-pass acceptance at the human gate (Tier 2) ___
- Edit depth per AI draft, trend (Tier 2) ___
- Sample-audit error rate on delegated work (Tier 2) ___
- Spam-flag rate vs human baseline (Tier 2) ___
- Pipeline from AI-assisted workflows, with baseline (Tier 3) ___
- One holdout comparison running at all times (Tier 3) ___

> **The holdout habit.** Whatever workflow you scale, keep a control: a content cluster written the old way, a territory without the auto-BDR, an email segment on the existing journey. Vendor case studies — including the ones quoted in this playbook — are existence proofs from other companies. Your holdout is the only evidence about yours.

The self-assessment

Twenty-four questions, four themes. Score each 0–3: **0** nobody has looked · **1** an opinion · **2** written down · **3** operating. ◆ marks the eight highest-signal questions. Same scale as the Growth Audit Framework, deliberately.

A System & context

___ / 18

01 ◆	Do your ICP, positioning, messaging and voice live in versioned files that every AI workflow reads?	0 1 2 3
02 ◆	Would a new hire running your best prompt get the same output quality as its author?	0 1 2 3
03	Is your voice file extracted from your published corpus, with exemplars — not the brand deck?	0 1 2 3
04	Has any qualification or scoring prompt been back-tested against accounts with known outcomes?	0 1 2 3
05	When positioning changed last, did downstream workflows update from one file — or by manual hunt?	0 1 2 3
06	Do outputs and results feed back into the context layer on a schedule?	0 1 2 3

B Workflows & human-in-the-loop

___ / 18

07 ◆	For each AI workflow: can you name where the human sits and what they are specifically checking?	0 1 2 3
08	Is delegation decided by verifiability and blast radius — written down — or by whoever felt like it?	0 1 2 3
09 ◆	Which outputs auto-ship without review, and who decided that on what evidence?	0 1 2 3
10	Do you run AI as critic — style pass, fact pass — before human review, on everything?	0 1 2 3
11	Do you have tiered rules — Tier-1 accounts human-reviewed, long tail auto — or one rule for all?	0 1 2 3
12	Is edit depth per draft tracked — and could you say whether its trend means quality or sleep?	0 1 2 3

C Governance & risk

___ / 18

13 ◆	Is there a written, marketing-specific AI policy — and was it audited against actual behaviour this year? 63% of organisations have none. A generic IT policy scores 1.	0 1 2 3
14	Could you list what customer or company data goes into which tools — and would security agree?	0 1 2 3
15 ◆	Do your incentives make people disclose their AI workflows, or hide them? Punitive policies create secret cyborgs.	0 1 2 3
16	Does a written fact-checking protocol name steps and owners for stats, product, competitor and pricing claims?	0 1 2 3
17	Who signs off when AI output touches pricing, legal claims or comparisons — by name?	0 1 2 3
18	Is your disclosure position on AI imagery, video and bylines a decision — or a default?	0 1 2 3

D Measurement & capability

___ / 18

19 ◆	Can you show pipeline or revenue impact from one AI workflow, with a before/after or holdout? Not time saved. Only 39% can. This is the question the rest exist for.	0 1 2 3
20	Is a holdout running right now on your most-scaled workflow?	0 1 2 3
21	Are quality metrics — first-pass acceptance, sample-audit errors, drift — on a dashboard anyone reviews?	0 1 2 3
22 ◆	Are you investing in the humans — training, review skill, taste — or only in tools? The market ratio is 5:1 tools over training. Backwards.	0 1 2 3
23	What are juniors no longer learning because AI does it — and is there a plan for that?	0 1 2 3
24	If your two heaviest AI users left this month, what capability walks out of the door with them?	0 1 2 3

READING YOUR TOTAL — / 72

Ad hoc — start at Part I	0–24
Emerging — fix the lowest theme	25–44
Operating — scale the library	45–60
Compounding — rare; verify	61–72

THE PATTERN TO EXPECT

Most teams score highest on theme B — individual workflows — and lowest on A and C, because tools were adopted before systems and rules. That profile is the definition of the 54% "ad hoc" cohort: real usage, no operating system. The 90-day plan overleaf is built for exactly that shape. As with Edition 01: score blind, compare with two colleagues, and treat a two-point spread on any question as an alignment finding.

The 90-day rollout

Three builds, one quarter, in an order that makes each build easier than the last. The discipline is what you refuse to add.

Wk 1–3

Build the context layer + declare amnesty

Week one: icp.md, positioning.md, messaging.md, proof.md, do-not.md — drafted from what exists, honest about what does not. Week two: extract voice.md from your ten best pieces. Week three: the amnesty audit — everyone demos how they actually use AI today, no penalties, and the six-part policy gets written from what you learn. You now know your real starting point, and every later build reads from these files.

Wk 4–8

Ship two workflows end to end

One quick proof (W5 repurposing or W6 reporting) and one revenue engine (W1 content or W2 account research), both with the full anatomy: AI critics, a named human gate, a checklist, and metrics from day one. Resist the third workflow — two running properly beats five running loosely, and the S&P Global abandonment data says loose pilots die.

Wk 9–12

Instrument, holdout, and harden

Stand up the eight-number dashboard. Start one holdout on the most-scaled workflow. Run the first voice-drift audit and the first 5% sample audit of delegated work. Fix what they find. End of quarter: re-score the 24 questions, blind, same people.

DO NOT, THIS QUARTER

- Buy a platform "to consolidate" before two workflows run
- Automate Tier-1 outbound
- Publish anything AI-drafted without the fact pass
- Chase a competitor's AI announcement
- Promise headcount savings to anyone

EVIDENCE YOU ARE DONE

- Context layer live, versioned, read by every workflow
- Two workflows in production with named gates
- Policy written from real usage, signed by the team
- Dashboard reviewed monthly; one holdout running
- Self-assessment re-scored, honestly

WHAT QUARTER TWO LOOKS LIKE

Add W4 to make the system compound, then W3 with volume caps once deliverability baselines exist. By then the question has changed from "should we use AI?" to "which constraint in our growth model does the next workflow relieve?" — which is Edition 01's question, and the two documents become one operating system.

Numbers you will hear that you should *not* repeat.

AI marketing has the worst statistics hygiene of any field this practice covers. The ones below circulate constantly; each is unverifiable, misread, or a prediction dressed as a measurement.

THE CLAIM	THE PROBLEM
"95% of AI pilots fail" (MIT)	The MIT NANDA report measured P&L attribution of custom enterprise pilots within months, from a small sample, and was not peer-reviewed. The same report found 90% of employees use AI personally with strong results. It is an argument about integration, not capability.
"10× productivity"	No primary source exists. The best controlled study — 758 professionals, BCG/Harvard — found +12% output, +25% speed, +40% quality. Large, real, and an order of magnitude short of the claim.
"Search volume will drop 25% by 2026"	A February 2024 Gartner *prediction*, almost universally cited as an observed fact. It did not materialise on schedule.
"30% of genAI projects abandoned"	Also a Gartner prediction (July 2024), not a measurement. The measured figure that exists: S&P Global found 42% of companies scrapped most AI initiatives in 2025.
"AI saves marketers 5 hours a week"	A 2023 Salesforce survey estimate by respondents — self-reported, vendor-fielded, endlessly recycled as measurement.
"342% ROI from [tool]"	Commissioned Forrester TEI studies model a hypothetical composite organisation, paid for by the vendor. A framework, not evidence.
"65% of marketing jobs won't survive AI"	Task-exposure modelling, not observed job losses. The observed data: creative-execution postings fell 20–33% year on year while strategy roles held — a reshaping signal, not a headline apocalypse.
Any vendor case study, uncritically	Including the ones quoted in this playbook. They are existence proofs from other companies' contexts. Your holdout is the only evidence about yours.

> **The habit, restated from Edition 01.** Before acting on a number: sample size, year, who paid for the research, and what was actually measured. In AI marketing, add a fifth check — *is it a measurement or a prediction?* The field's most-quoted figures fail that one.

Sources

SURVEYS & BENCHMARKS

Content Marketing Institute / MarketingProfs, *B2B Content Marketing Benchmarks* (2025, n=980; 2026, n=1,015)

ICONIQ Growth, *The State of GTM* (2025, n=194; 2026, n≈150)

Gartner, *CMO Spend Survey* (2025, 2026, n=401) and consumer AI-search survey (2025)

McKinsey, *The State of AI* (2025, n=1,993)

KeyBanc / Sapphire, *Private SaaS Company Survey*, 16th annual (2025)

High Alpha / Kyle Poyar, *2025 SaaS Benchmarks* (n=800+)

S&P Global Market Intelligence, AI initiative survey (2025)

Microsoft / LinkedIn, *Work Trend Index* (2024, n=31,000)

Lightcast, AI skills premium analysis, 1.3B+ postings (2025)

Bloomberry / Revealera, 180M job postings analysis (2025)

HubSpot, *State of AI* (2025); Salesforce, *State of Marketing* (2026, n=4,450) — both vendor-fielded; read their figures with that in mind

SEARCH & CONTENT EVIDENCE

Pew Research Center, AI-summary click study (2025, 68,879 searches)

Ahrefs, AI Overviews CTR study (2025, 300K keywords)

Semrush, AI vs human ranking study (2025–26, 42K posts); survey of SEO teams (2025, n=224)

Microsoft Clarity via Digiday, LLM referral conversion (2025)

Similarweb, genAI traffic and citation data (2025–26)

Google Search Central, March 2024 core update & spam policies

WORKFLOW DOCUMENTATION

Kyle Poyar, *Growth Unhinged* — the AI-for-GTM teardown series: the modern ABM engine (Dan Rosenthal), AI agents for marketing (SafetyCulture), the AI GTM system (Matteo Tittarelli), AI-native GTM plays, the AI-native growth team (Fyxer), public-company scale (monday.com), *2026 State of AI GTM Report* (with Maja Voje)

Animalz — *AI Content Works (But Only If You Do the Work)*; *Don't Stop at Workflows: Build a Compounding Content System*; Claude Code style-guide enforcement (via The Workflow)

HubSpot editorial team, *How the HubSpot Blog Team Uses AI* (2025)

HUMAN-AI COLLABORATION RESEARCH

Ethan Mollick, *Centaurs and Cyborgs on the Jagged Frontier*; *Detecting the Secret Cyborgs* — oneusefulthing.org

Dell'Acqua et al. (BCG / Harvard), *Navigating the Jagged Technological Frontier* (2023, n=758)

Lee et al. (Microsoft / CMU), *The Impact of Generative AI on Critical Thinking*, CHI 2025

GOVERNANCE & LEGAL

US Copyright Office, *Copyright and Artificial Intelligence, Part 2* (2025)

European Commission, AI Act Article 50 transparency FAQ (application August 2026)

Moffatt v. Air Canada, BC Civil Resolution Tribunal (2024)

Smart Insights (Dave Chaffey), AI marketing policy example (2025); MarTech.org, the AI oversight gap (2025)

Vectara Hallucination Leaderboard (open benchmark, 2024–26)

Anthropic, prompt-engineering best practices (2025)

Vendor-published data — HubSpot, Salesforce, Semrush, Ahrefs, Microsoft — is used where it is the best available and flagged as such where it appears. Where a claim rests on a single company's teardown, it is presented as an existence proof, not a benchmark. Predictions are labelled as predictions.

Thirty minutes on your *AI setup.*

If you have scored the self-assessment — or you just know your team's AI usage grew faster than its rules — book a free strategy call. Thirty minutes, no commitment. We will find the one build from this playbook that moves your quarter, and I will tell you plainly if you do not need help doing it.

Bring your four theme scores if you have them. No deck, no proposal, no nurture sequence — the same terms as everything else this practice publishes.

BOOK THE CALL

partneringrowth.co.uk

Free 30-minute strategy call — no commitment

OR ON LINKEDIN

Hilal Tasdan

Message "playbook" with your four scores

PARTNER IN GROWTH

Fractional CMO and go-to-market work for B2B SaaS companies scaling pipeline in crowded markets. London, working across the UK and Europe. Embedded leadership rather than a retainer — in your standups, your Slack and your CRM.

Also in this series: *The SaaS Growth Audit Framework* (Edition 01) — the diagnostic this playbook's workflows plug into.

SELECTED RESULTS

Access PaySuite · TOFU growth	54×
Checkbook.io · paid CAC	-41%
Uptechlab · qualified leads	3.1×
Client retention	94%