

AI PRODUCTIVITY IN B2B MARKETING

Does AI Actually Save Marketing Time?

What controlled studies show - and what B2B teams should measure before claiming productivity gains.

The answer: often, but not automatically.

The strongest studies show faster or higher-quality output on clearly bounded tasks.

They also show performance can fall when a task sits outside the model's capability frontier.

This report reviews four controlled or quasi-experimental studies across writing, customer support, consulting and software development.

Executive findings

The research does not support one universal “AI saves X%” claim. It supports a more useful conclusion: productivity depends on the task, the worker’s starting point, and how quality is evaluated.

1. Clear, repeatable work can improve materially

In a randomized professional-writing experiment, ChatGPT reduced completion time by about 40% and raised rated output quality by 18%.¹ Customer-support agents with AI assistance resolved 15% more issues per hour on average.²

2. Quality can improve - until the task exceeds the tool

In a consultant experiment, participants using GPT-4 completed 12.2% more tasks, finished 25.1% faster and produced more than 40% higher quality on tasks inside AI’s capability frontier. On outside-frontier tasks, they were 19 percentage points less likely to reach the correct answer.³

3. The unit that matters is accepted work

A quick draft is not the same as a useful deliverable. Marketing teams should measure the time from brief to approved output, then record first-pass acceptance, rework and factual or brand issues.

Context: adoption is ahead of proven business impact

Survey results are useful context, but they are not a substitute for task-level measurement.

80%

of respondents in McKinsey’s 2026 survey reported that AI had improved individual productivity. This is self-reported, not a timed workflow study.

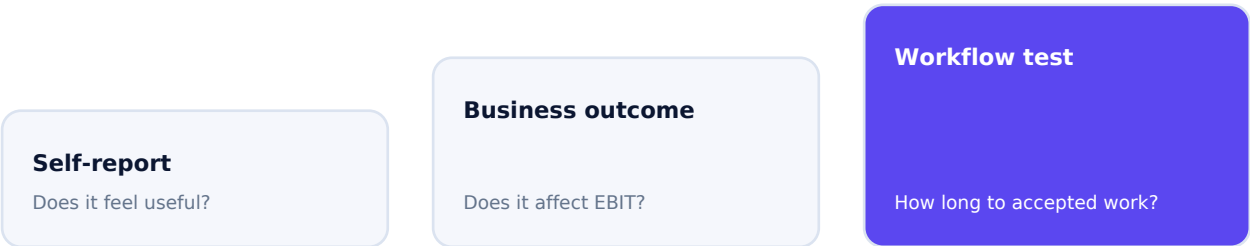
37%

reported at least some EBIT impact from AI use. Business effects appear less widespread than individual productivity effects.

20,000

knowledge workers were surveyed by Microsoft for its 2026 Work Trend Index. Its finding is relevant here: people expect to remain accountable for oversight, judgment and quality control.

What different evidence can and cannot show



This report uses all three forms of evidence, while prioritizing workflow-level measures.

Evidence reviewed

Inclusion rule: the study had to compare AI access with a baseline and report at least one observable output, time or quality outcome. Results below are not pooled because the tasks, measures and workers differ.

Study	Setting	Design / sample	Measured result
Noy & Zhang (2023) ¹	Professional writing tasks	Randomized experiment	Time -40%; rated quality +18%
Brynjolfsson, Li & Raymond (2023) ²	Customer support	Staggered rollout; 5,172 agents	Issues resolved per hour +15% on average
Dell’Acqua et al. (2023) ³	Consulting tasks	Experiment; 758 consultants	Inside frontier: 12.2% more tasks; 25.1% faster; >40% higher quality
Peng et al. (2023) ⁴	Software development	Controlled experiment	HTTP-server task completed 55.8% faster

The evidence is directional for marketing, not a direct marketing benchmark.

Writing is closest to many content tasks. Support, consulting and coding expand the evidence base, but their exact percentages should not be transferred into a B2B content forecast.

Results are strong - and highly task-specific

Reported gains from four studies, shown separately because the underlying measures differ.

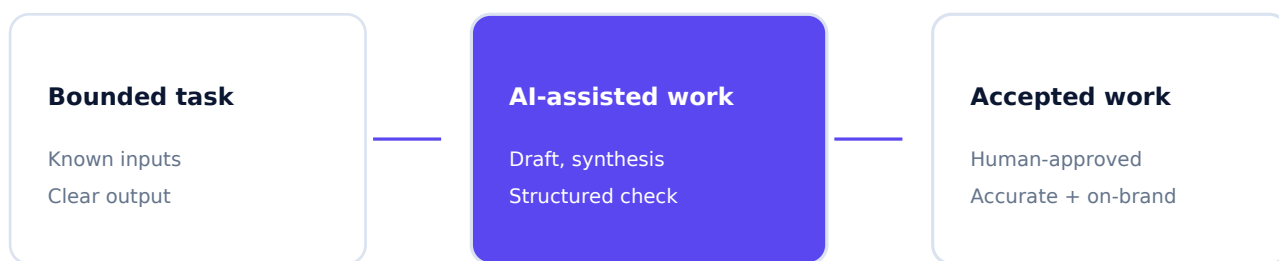


Do not average these numbers.

They measure different work, populations and outcomes. An “average AI productivity gain” would look precise but would not be meaningful. Compare your own recurring task and quality gate.

Where the AI-first story breaks

The most important caution in the reviewed evidence comes from the consulting experiment. On tasks beyond AI's capability frontier, consultants with AI were 19 percentage points less likely to produce a correct solution.³



For marketing, the risk rises when output makes unverified factual claims, interprets business context incorrectly, or must meet a nuanced brand standard. These are review design problems as much as model problems.

Teamulate evidence synthesis

We coded the four studies against five questions that matter for B2B marketing operations. This is an original analytical framework, not a new human-subject experiment.

Question	Synthesis
Is there a measured speed gain?	Yes. All four studies report a speed or throughput gain on their tested work.
Does quality always fall?	No. Two studies report higher quality on their evaluated tasks. Quality can decline when the task exceeds the model's frontier.
Can one percentage be used for marketing?	No. Outcomes are heterogeneous and the evidence is not a direct test of end-to-end B2B marketing production.
Who benefits most?	Several studies report larger gains for less experienced or lower-performing workers. This is a pattern, not a guarantee for every team.
What should a team measure?	Human minutes to accepted work; first-pass acceptance; rework; factual and brand issues; variable tool cost.

Analytical conclusion

The most defensible gains happen when teams define the task boundary, supply context and retain a quality gate.

A measurement model for B2B marketing teams

Use this model to create your own valid benchmark. Compare one workflow before and after a clearly documented AI intervention; do not compare two unrelated tasks.

Metric	Definition	Decision it supports
Human minutes to accepted work	Hands-on time from intake through review and approval.	Capacity / resourcing
First-pass acceptance	Share of first submissions that meet the acceptance checklist.	Quality of the workflow
Rework cycles	Material revisions after first submission, with reason logged.	Where process design fails
Factual / brand issues	Source, accuracy, audience-fit and tone issues found in review.	Risk and quality control
Variable cost per accepted asset	Direct tool usage cost divided by approved deliverables.	Unit economics

Minimum viable pilot

Select one recurring asset type. Run 10 to 15 comparable briefs under the current workflow, then repeat with AI assistance and the same acceptance checklist. Record the model, source pack, reviewer and every material revision. Report the full result, including no change or decline.

How to validate this in your own team

The studies establish that task design and quality gates matter. The next step is a comparable internal benchmark that captures the full workflow, from a written brief to an approved asset.

Benchmark question

When briefing, source checking, quality review, correction and approval are included, how much human time does AI save on recurring B2B marketing work?

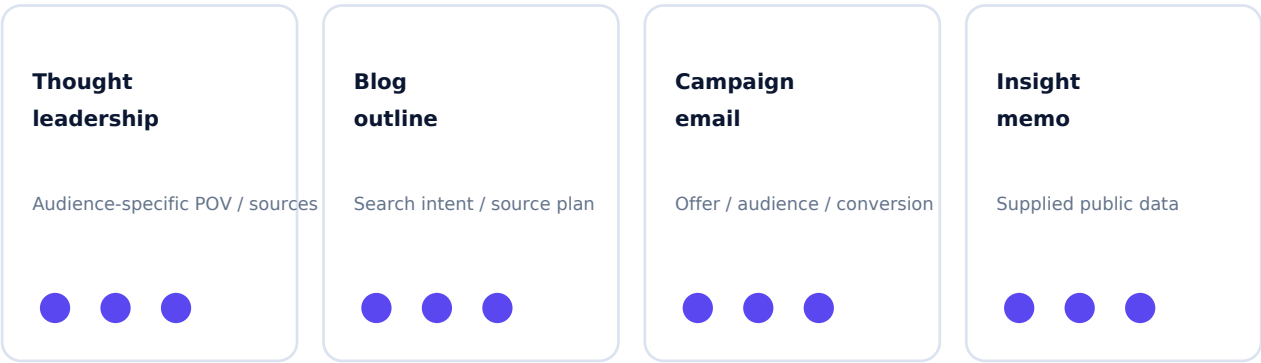
Condition	What is tested	What stays constant
1. Human-only	Work completed without generative AI.	Brief, factual source pack, quality bar and reviewer.
2. One AI assistant	One operator uses a general AI assistant and remains responsible for the result.	Same brief, sources, quality bar and reviewer.
3. Coordinated workflow	AI-supported production with documented handoffs, review checklist and human approval.	Same brief, sources, quality bar and reviewer.

Principle: same input. Same quality bar. Different ways of working.

That is what makes differences in time, quality and rework comparable.

A 12-brief set for B2B marketing

Four common asset types, each repeated across three comparable briefs.



Brief requirements

Each brief should contain the same elements: audience, objective, named source pack, required claims, brand constraints, deliverable format and acceptance checklist. Avoid tasks that require open-ended external research during the benchmark.

Why three briefs per task?

One task can be unusually easy or difficult. Three variants reduce the risk that a single brief drives the result. A small benchmark should report task-level results alongside any category summary.

Measure accepted work, not draft speed



Metric	Definition	Why it matters
Active human minutes	Hands-on time for briefing, creation, review, correction and approval.	Capacity consumed.
Time to acceptance	Elapsed time until the work meets the agreed criteria.	Whole-workflow speed.
First-pass acceptance	Initial submission passes review without material revision.	Quality, not output volume.
Rework cycles	Count and reason for material changes after submission.	Hidden operating cost.
Factual and brand issues	Logged source, accuracy, audience-fit and tone issues.	Risk visibility.
Variable tool cost	Direct per-task AI usage cost, where applicable.	Cost per accepted asset.

Benchmark controls and reporting rules

A credible internal result requires controls. These rules prevent the benchmark from becoming a demonstration designed to confirm a pre-chosen answer.

Balance the order

Rotate the order of conditions to reduce learning and fatigue effects.

Record the context

Log operator, model/version, source pack, prompt changes, timing and review outcome.

Set acceptance first

Define acceptance criteria before the benchmark begins; a fast output that fails review is not accepted work.

Use the same reviewer

Keep the reviewer and checklist consistent across comparable work.

Publish the whole result

Report no change and declines as well as positive gains. Show task-level variation.

State the boundary

Describe the tasks, operators and tools tested. Do not generalize results to every team or AI system.

Minimum viable benchmark

Run 10-15 comparable briefs per condition when feasible, document every material revision, and calculate results only after every brief has been reviewed.

Sources and method notes

This report combines an evidence review with an operational benchmark framework for B2B marketing teams. Teamulate did not conduct the external experiments cited below. The benchmark framework is a recommended method for teams to test their own work; it does not report Teamulate performance results.

Sources

1. Noy, S. and Zhang, W. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381, 187-192 (2023). <https://doi.org/10.1126/science.adh2586>
2. Brynjolfsson, E., Li, D. and Raymond, L. "Generative AI at Work." NBER Working Paper 31161 / arXiv (2023). <https://arxiv.org/abs/2304.11771>
3. Dell'Acqua, F. et al. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." SSRN (2023). <https://doi.org/10.2139/ssrn.4573321>
4. Peng, S. et al. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot." arXiv (2023). <https://arxiv.org/abs/2302.06590>
5. McKinsey. "The State of AI in 2026." 25 Aug 2026. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2026>
6. Microsoft. "2026 Work Trend Index." 5 May 2026. <https://www.microsoft.com/en-us/worklab/work-trend-index>

Interpretation rules

The report does not claim that any published result will replicate in every marketing team. The cited studies differ in task, setting, population, model and quality measure. Percentages are reported as published and are not combined into a single average.

About Teamulate

Teamulate is a managed AI marketing department for growing B2B companies. It combines AI execution with human judgment to help teams create, review and improve marketing work without building a large in-house department.