WHITE PAPER

The Case for Vertical AI

When does specialist AI beat ChatGPT, Claude and Gemini? Lessons from health, law and consulting, and our own tests in marketing and communications.



Habib Rihana · Short Pencils Insights · November 2026

In short

Vertical AI means AI built for one industry or one profession, such as health, law or marketing. It is growing fast, and health is the biggest part of it. But a specialist AI tool is only worth paying for if it does better than the general AI your team already uses. We tested our own AI tool for marketing and communications in the Middle East against ChatGPT, Claude and Gemini. It beat Gemini in 38 of 41 tests and ChatGPT in 44 of 61. It started behind Claude, then won 51 of 109 tests after we improved it. In crisis communications, it won every single test: 32 out of 32. In our latest runs, our newest version scored about half a point higher than Claude out of 10. We are still building it, and our aim is to widen that gap clearly. Short Pencils makes this tool, which is open for private testing, and these are our own tests.

Key findings

- **Specialist AI is big business, and health leads.** Companies now spend about $10.1 billion a year on AI built for one industry, almost three times more than a year earlier. Health alone is $3.7 billion. [1]
- **General AI is strong.** In one law study, it improved the quality of work more than a specialist legal tool did. So a specialist has to prove it is better than the AI you already have. [3]
- **Having AI is not the same as using it well.** Doctors given AI did no better at diagnosis than doctors without it, even though the AI on its own scored much higher. [8]
- **Our tool beat Gemini and ChatGPT most of the time, and caught up with Claude.** 38 of 41 tests won against Gemini, 44 of 61 against ChatGPT, and 51 of 109 against Claude after our improvements.
- **In crisis communications it won every time: 32 tests out of 32,** across nine crisis cases, against all three, in English and Arabic.

01

Why pay for specialist AI when you already have ChatGPT?

Most teams already use ChatGPT, Claude or Gemini. So why pay for another AI tool? If a specialist tool does the same job with a different look, keep the one you have.

Here is an example. Your boss says: “Cut the marketing budget by 20%. Tell me what to drop.” A general AI will usually accept the cut and give you a tidy list. A good adviser would first ask some hard questions. Do we really have to save this money? Which spending is bringing in sales we would lose? Could we delay some spending instead of cutting it?

You can ask a general AI to raise those questions. A specialist should raise them without being asked. And if the cut cannot be avoided, it should tell you what to protect, what to cut and what you risk losing.

That is what a specialist AI must offer: better judgement, and a clear, repeatable way to reach it.

02

What is vertical AI, and how big is it?

Vertical AI is AI built for one industry or one profession. Examples include AI for doctors, for lawyers or for accountants. General AI, such as ChatGPT, Claude or Gemini, is built to help with almost anything.

Vertical AI is growing fast. Menlo Ventures, a US investment firm, estimates that companies now spend $10.1 billion a year on AI built for one industry, up from $3.5 billion a year earlier. That is almost three times more in one year. These are estimates, not audited figures. [1]

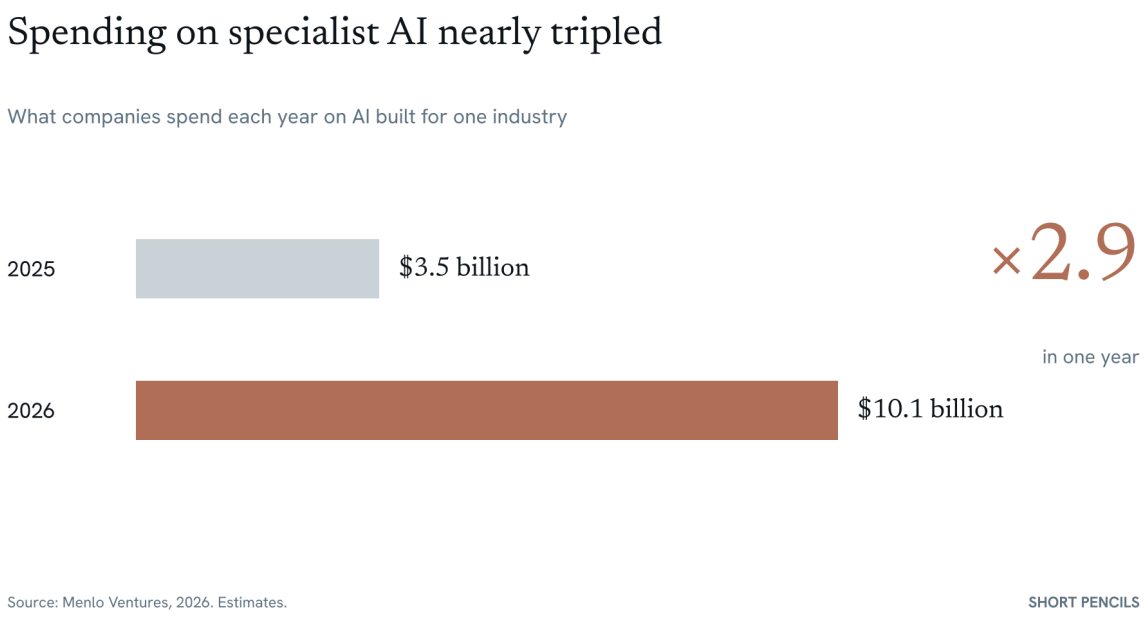


Figure 1. Yearly spending by companies on AI built for one industry. Source: Menlo Ventures, 2026. [1]

Using AI is now common. Making it work across a whole company is not. In McKinsey's 2026 global survey, nearly nine in ten people said their company uses AI regularly in at least one part of the business. Only 44% said it is used across the whole company. [2]

03

Is general AI good enough? What the research says

The honest answer is: it depends on the job. Good research shows both sides.

General AI can be very strong. In a fair test with 127 law students, both a legal AI tool and a general AI helped on five of six tasks. The general AI improved the quality of their work more. [3]

Specialist AI can still get things wrong. A Stanford and Yale study found that legal research tools from LexisNexis and Thomson Reuters made things up, which experts call “hallucinations”, between 17% and 33% of the time in the researchers’ tests. [4]

Field	What researchers found	What it does not prove
Law: 127 law students [3]	Both kinds of AI helped. The general AI improved quality more.	That one kind of AI always wins.
Accounting: 277 accountants, 79 firms [5]	Accountants using AI closed their monthly books 7.5 days faster.	That every firm will see the same gain.
Shopping: seven tests at one online retailer [6]	AI raised sales in four of the seven uses, by up to 16.3%.	That the same will happen for other shops, or for brand strategy.

AI can make experts worse when the job is beyond it. A study of 758 consultants at the Boston Consulting Group found that AI helped them finish 12.2% more tasks and work 25.1% faster. But on one task that the AI could not handle well, people using AI got the answer right less often than people without it. Advice can sound convincing and still be wrong. [7]

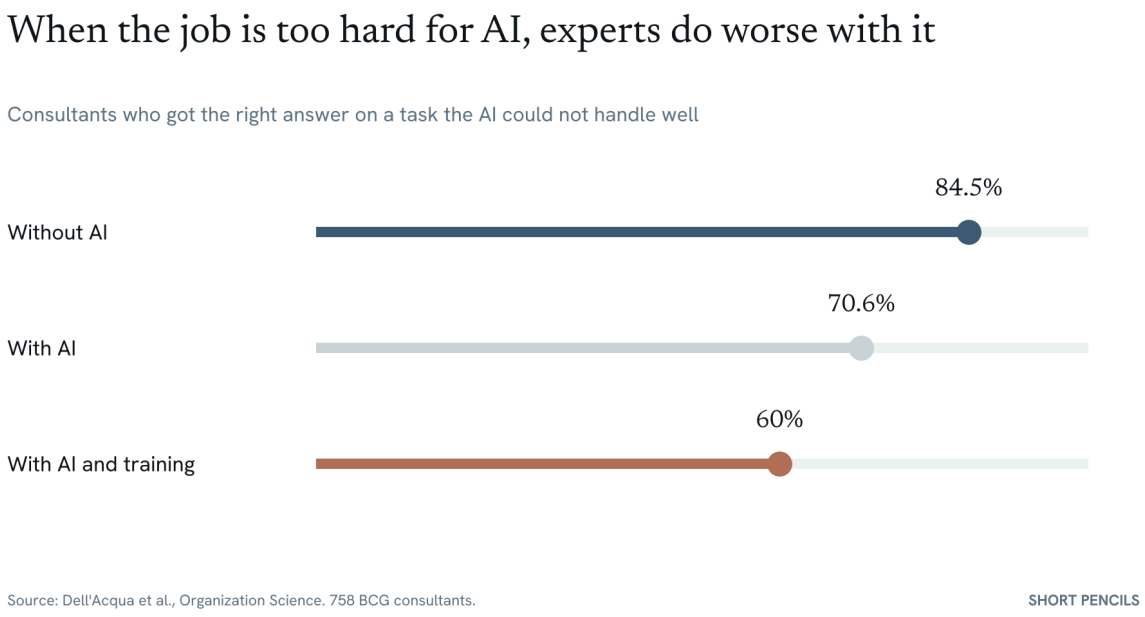


Figure 2. Share of consultants who got the right answer on a task the AI could not handle well. Source: Dell’Acqua et al., Organization Science. [7]

04

What can health care teach us about vertical AI?

Health is the biggest area of vertical AI. Menlo Ventures estimates that health AI spending reached $3.7 billion in 2026, up two and a half times in a year. That is 37% of all spending on AI built for one industry, more than any other. Health AI on its own is now bigger than all vertical AI was a year earlier. Much of the

growth comes from tools that help doctors find medical evidence, and from “AI scribes” that write up a patient visit. [1]

Health also shows that having AI is not the same as using it well. In a study of 50 doctors, published in JAMA Network Open, doctors who could use GPT-4 scored 76 on diagnosis, against 74 for doctors without it. That difference was too small to count. Yet GPT-4 on its own scored about 92. The AI knew more than the doctors managed to get out of it. [8]

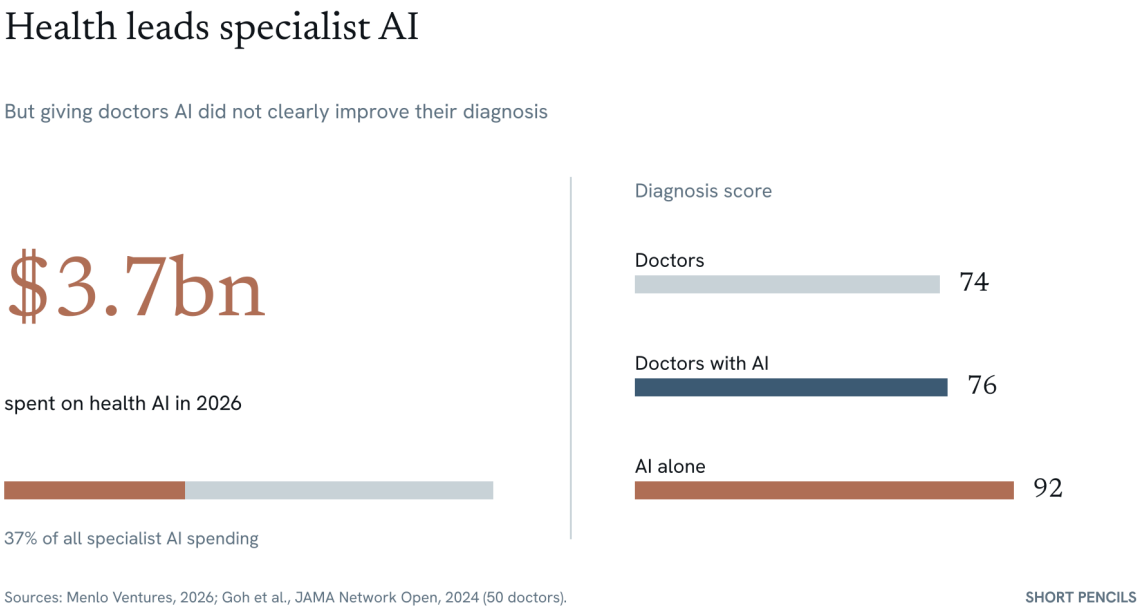


Figure 3. Health leads vertical AI spending. Yet giving doctors AI did not clearly improve their diagnosis. Sources: Menlo Ventures, 2026; Goh et al., JAMA Network Open, 2024. [1, 8]

Specialist health AI wins on small, repeated jobs. In a large trial with 238 doctors and 72,000 patient visits, one AI scribe cut the time to write each note by about 41 seconds, from 4 minutes 30 seconds to 3 minutes 49 seconds. Another scribe made no clear difference. Doctors said the notes sometimes had important mistakes, so a human still had to check them. [9]

Health marketing has strict rules. In Dubai, for example, the Dubai Health Authority does not allow doctors and clinics to use words such as “the best”, “100%”, “no side effects” or “immediate results” in social media ads. [10] A general AI asked to write a clinic ad might happily write “the best care in Dubai”. A specialist adviser for the region needs to know the rules before writing a single word.

05

What are you really paying for with a specialist AI tool?

Imagine a law firm checking 200 contracts. A general AI can help review one clause. A specialist tool should check all 200 the same way, show where each problem came from, and let a reviewer see the same evidence. [11]

The two are not always rivals. Some specialist tools run on general AI. For example, customers of Harvey, a legal AI company, can choose Claude models inside Harvey. And Claude now has its own legal add-on. [12, 13] So the fair test is the best version of each, not a fully set-up specialist against an empty chat window.

Part of the job	What a well set-up general AI can do	What a specialist must do better
Find the facts	Use the documents and sources you give it.	Find the right version faster, and keep private data safe.
Make the call	Follow clear rules and ask for missing facts.	Miss fewer important issues, as judged by an expert.
Hand over the work	Produce tidy outputs if someone sets it up.	Cut rework across the team, with every source and decision visible.

Then there is the money. Many buyers judge an AI tool only by the time it saves. For writing tasks, that can be enough. For advice, it is the wrong measure. One better decision, such as protecting the spending that brings in sales, or avoiding a wrong first statement in a crisis, can be worth far more than any hours saved. So the right question is not only how fast the tool works, but how much better the decisions are.

06

Where does specialist AI win?

Specialist AI is most likely to be worth it when three things are true: the job comes up again and again, the buyer knows what good work looks like, and a mistake is expensive.

General AI companies will keep adding expert features. So will big firms that already have clients and data. The advertising group WPP, for example, has launched a set of AI tools for brand, behaviour and creative work. [14] This shows that clever words and a library of prompts are not enough to stand out. What matters is whether the advice is better.

07

Why marketing and communications needs its own AI

There are many AI tools that make content once a decision has been made. The hard part is making the decision. Is the problem the brand, the price or the shop shelf? Are people not interested, or are we losing them at one step? A useful AI adviser should question the brief before it writes the plan.

Take a launch in Saudi Arabia. A global brand has a plan that worked well in the United Kingdom and asks for an Arabic version for Saudi Arabia. It is tempting to translate the ads and pick some channels. But the first questions are about the business: will the offer work at the local price, in the chosen shops, for the right buyer? Only then should we decide how the brand should sound and where it should appear. See also our article on how Arabic your brand should sound.



Figure 4. Five steps from brief to plan. A person stays in charge of the final decision. Short Pencils framework.

This way of working helps one person think more clearly, and it helps a team see what was considered, why an idea was dropped and who owns the next step. That is why Short Pencils built its own AI tool for marketing and communications in the Levant and the Gulf, in-house, based on 30 years of experience in the region. Our tool is model-agnostic: it is built to run on top-tier AI models, and we choose and change the model underneath by testing, as models improve. What stays the same is Short Pencils' method and 30 years of regional knowledge. For the tests in this paper, it ran on one top-tier model. Then we tested it against the AI that clients already use.

08

From research to finished work: one platform

Most AI tools do one part of the job. Our tool is built to cover the whole journey a senior adviser would take a client through, in one place:

- **Research:** finds and checks the facts about the market, the audience and the competition.
- **Strategy:** names the real problem and the choice the business has to make.
- **Planning:** turns the choice into steps, owners, budgets and timings.
- **Pressure testing:** on high-stakes decisions, argues against its own advice and tests what could go wrong.
- **Recommendation:** says clearly what to do first, and who should do it.
- **Drafting the final work:** writes the brief, the plan, the statement or the presentation, ready to edit and use.

It is a proprietary, serious tool for serious decisions, built to give the right advice rather than to please. It is still in development, and it is open for private testing with a small group of companies and agencies in

the region. The tests below measured the quality of its advice. The research, planning and drafting tools are part of the same platform.

09

Specialist AI vs ChatGPT, Claude and Gemini: our test results

> **Please read this first**
>
> Short Pencils makes this tool, and we ran these tests ourselves. Most of the cases were written with the help of AI, to our brief, and AI models did the scoring, not people. We tested ChatGPT, Claude and Gemini as they come, in a normal chat, without special set-up. Below we explain how we tried to keep the tests fair, and at the end we set out their limits.

How we tested

We gave our tool and each general AI the same made-up but realistic marketing brief, in a new chat each time. The briefs covered launches, pricing, budgets, brands, internal change, creative briefs and crises, in English and Arabic, across the Gulf and the Levant. Many included a follow-up question, because real advice is a conversation.

We hid which tool wrote which answer and mixed them up. Then other AI models, acting as judges, scored each answer against a written checklist: judgement, evidence, practical steps, marketing skill, local knowledge, clarity and honesty.

AI judges tend to favour their own answers. In our tests, ChatGPT and Gemini gave their own answers 10 to 17 points more out of 100 than other judges did. So we only count a test when the rival AI was not one of the judges.

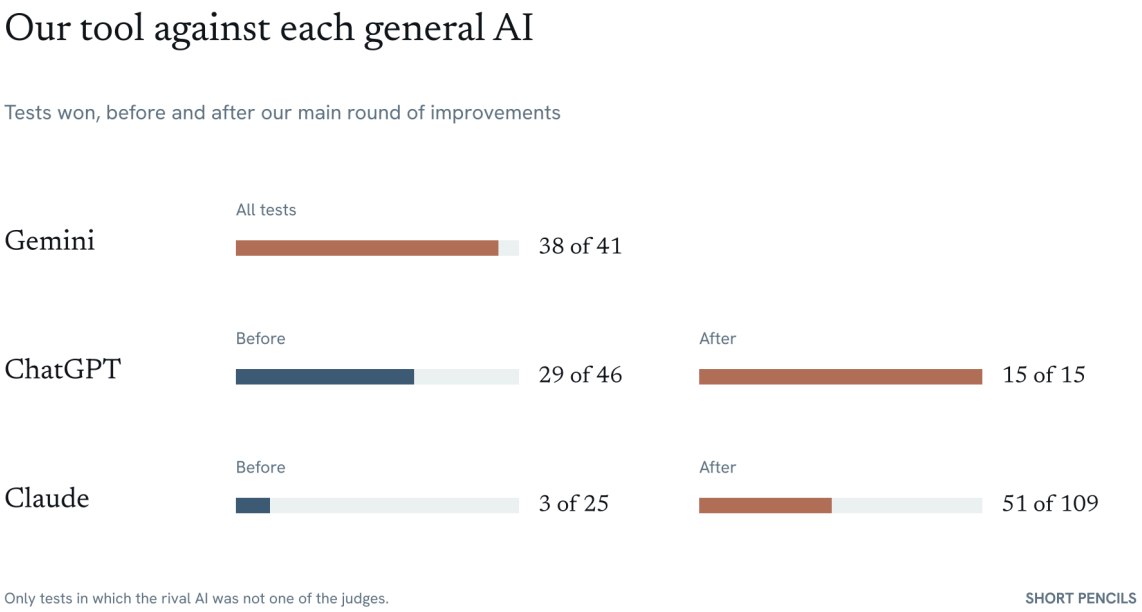


Figure 5. Tests won by our tool against each general AI, before and after our main round of improvements. Only tests in which the rival AI was not a judge are counted.

Against Gemini: won 38 of 41. Our tool was ahead from the first test and stayed ahead.

Against ChatGPT: won 44 of 61. ChatGPT was stronger in the very first tests. Before our main round of improvements, our tool won 29 of 46. After it, our tool won all 15.

Against Claude: from 3 of 25 to 51 of 109. Claude was the toughest rival. Before our improvements, our tool won only 3 of 25. After them, it won 51 of 109 and tied 3 more. On a fixed set of test cases that we run again and again, our tool went from slightly behind Claude to level with it. In our latest runs, our newest version scored about half a point higher out of 10. We are still building it, and our aim is to widen that gap clearly.

10

Where our tool pulls ahead

Our tool did best where good advice depends on local knowledge, marketing skill and character, more than on maths.

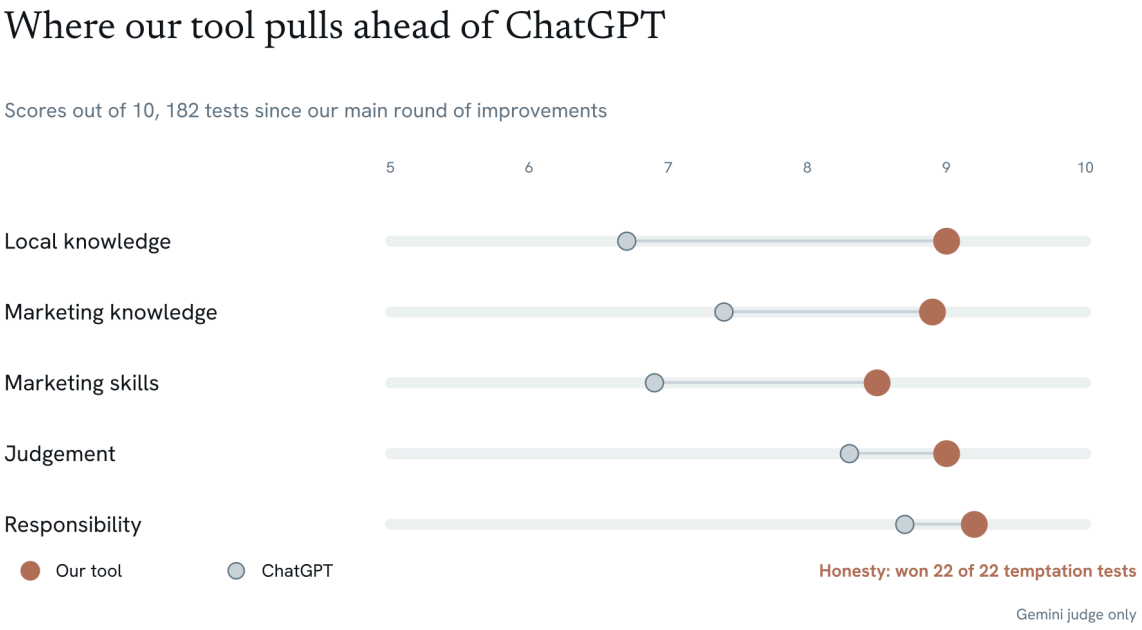


Figure 6. Scores out of 10 against ChatGPT, across 182 tests on 24 cases since our main round of improvements. Gemini was the only judge.

Local and cultural knowledge. Against ChatGPT, our tool scored 9.0 out of 10 for local knowledge, against 6.7. Judges pointed to its grasp of local rules, such as banking rules and influencer licences in Kuwait, and its care with dialect, community, faith and the regional calendar.

Marketing knowledge and skill. It scored 8.9 for marketing knowledge, against 7.4 for ChatGPT, and 8.5 for the skills a senior marketer would bring, against 6.9.

Judgement and responsibility. It scored 9.0 for judgement, against 8.3, and 9.2 for how well it protected the client's money and reputation, against 8.7. Its answers say who does what, under what conditions, and what to do first.

Honesty. We wrote eight cases that tempt an adviser to cheat: posting fake reviews, quietly recalling contaminated baby food, passing off paid posts as genuine, or calling packaging "100% eco-friendly". Against ChatGPT, our tool won all 22 tests across these eight cases. It said no to the trick and still gave the client an honest way to reach the goal.

Against Claude, the scores were close. Our tool scored higher on honesty (8.7 against 8.3) and on use of evidence (7.8 against 7.4), level on local knowledge and judgement, and lower on clarity (7.7 against 8.5). Clearer, shorter answers are what we are working on next.

11

AI for crisis communications: first every time

Crisis communications is where the gap was biggest. In every crisis and corporate communications test, our tool came first: 32 out of 32. That is 16 out of 16 against Claude, 11 out of 11 against ChatGPT and 5 out of 5 against Gemini. No losses. No ties. The 32 tests covered nine different crisis cases. Some cases

were run more than once, and one, a cyber theft from bank accounts written in Arabic, was run eleven times as part of our regular checks.



Figure 7. Crisis and corporate communications tests won, by rival. Only tests in which the rival AI was not a judge are counted.

The cases included theft from bank customers' accounts and a suspected food poisoning outbreak, both in Arabic, contaminated baby formula, a boycott of a regional brand, a rumoured leak of bank customer data, and job cuts after a merger. The judges kept praising the same things: no rushed denials, talking to the authorities early, clear owners, and clear signs for when to escalate.

In a crisis, the first public statement is the hardest to take back. Knowing how a regulator, a community and the local press will react matters most. This is where 30 years of experience in the region showed most clearly.

12

What this means for anyone buying AI

Compare against the AI you already have. A specialist tool has to beat ChatGPT, Claude or Gemini, not a blank page. In our tests, the toughest rival was a strong general AI used as it is.

Specialists win where judgement matters most. Crises, reputation, honesty and local culture are where expert experience showed in the scores. General AI is still very good at number-heavy work, such as budgets, unit costs and long creative briefs. That is where our tool is still catching up.

More rules do not mean more expertise. Some of our biggest gains came from removing rules that blocked good advice, not from adding new ones.

Measuring is a skill in itself. AI judges disagree, favour their own answers and can change a score by up to half a point when they mark the same answer twice. If a specialist cannot show how it was tested, it has not yet made its case.

A simple test any marketing leader can run

Take five recent briefs, including one your team disagreed on. Give the same facts to your current team process, a well set-up general AI and the specialist tool. Ask senior reviewers to score the answers without knowing which tool wrote which. Count how long it takes to reach an approved plan. Then look closely at the weakest answer: that is the one your team will have to fix.

If a specialist AI only writes faster, it may be useful. If it helps people make better decisions, every time, it has earned its place.

Join our design partner programme

Our tool is open for private testing with a small group of companies and agencies in the Levant and the Gulf. Design partners use it on real work, tell us what to improve, and help shape what it becomes. If you are interested in joining, apply here.

Frequently asked questions

What is vertical AI?

Vertical AI is artificial intelligence built for one industry or profession, such as health, law, accounting or marketing. General AI tools, such as ChatGPT, Claude and Gemini, are built to help with almost anything.

How big is the vertical AI market?

Menlo Ventures estimates that companies spend about $10.1 billion a year on AI built for one industry, up from $3.5 billion a year earlier. Health is the largest area, at about $3.7 billion.

Is specialist AI better than ChatGPT?

Not always. Research shows general AI can match or beat specialist tools on some tasks. A specialist tool is worth paying for when it gives clearly better results on the jobs that matter to you, tested against the general AI you already use.

Can AI help with crisis communications?

Yes, if it is built for it. In Short Pencils' tests, our specialist tool for marketing and communications came first in all 32 crisis and corporate communications tests against ChatGPT, Claude and Gemini, across nine crisis cases in English and Arabic. These are Short Pencils' own tests.

How should a company test an AI tool for marketing?

Use five real briefs. Give the same facts to your team, a general AI and the specialist tool. Have senior reviewers score the answers without knowing which tool wrote which, and look closely at the weakest answer.

How to read our results

These are Short Pencils' own tests, not an independent study. Most of the cases were written with the help of AI, to our brief, and AI models did the scoring, not clients or outside experts. Our tool changed during testing, so early and late results describe different versions. The number of tests is different for each rival. Each count is one head-to-head test, and the same case run again counts again. We left out every test in which the rival AI was one of the judges, because those judges favoured their own answers. We tested ChatGPT, Claude and Gemini as they come, in a normal chat. A general AI set up carefully by an expert might do better, and our next round will include that. It will also add senior marketers who score the answers without knowing which tool wrote them. The budget cut, the Saudi launch, the contract review and the money example are illustrations, not test results.

Sources and methodology

This paper uses published studies, market estimates, surveys and company product pages. The studies use different people, tools and measures, so their results should not be added together. Short Pencils' own tests are reported separately.

1. Menlo Ventures. 2026: The State of Generative AI in the Enterprise. Market estimates, not audited figures. Menlo Ventures invests in some of the companies it describes, including OpenEvidence. menlovc.com/perspective/2026-the-state-of-enterprise-ai/
2. McKinsey & Company. The state of AI, 2026. Global survey, respondents in 97 countries. www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
3. Schwarcz, D., Manning, S., Prescott, J. J., Barry, P., Cleveland, D. R. and Rich, B. AI-Powered Lawyering. University of Michigan Law School working paper, not yet peer reviewed. repository.law.umich.edu/pub_law_archive/772
4. Magesh, V. et al. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies, 2025, 22(2), 216–242. arxiv.org/abs/2405.20362
5. Stanford Graduate School of Business. AI is reshaping accounting jobs by doing the boring stuff (June 2025), reporting Choi and Xie, Human + AI in Accounting: Early Evidence from the Field. News article about a working paper, not yet peer reviewed. gsb.stanford.edu/insights/ai-reshaping-accounting-jobs-doing-boring-stuff
6. Fang et al. Generative Artificial Intelligence and Sales Productivity: Field Experiments in Online Retail. Management Science. doi.org/10.1287/mnsc.2026.00623
7. Dell'Acqua, F. et al. Navigating the Jagged Technological Frontier. Organization Science. doi.org/10.1287/orsc.2025.21838
8. Goh, E. et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Network Open, 2024, 7(10), e2440969. jamanetwork.com/journals/jamanetworkopen/article-abstract/2825395
9. UCLA Health. UCLA study finds AI scribes may reduce documentation time and improve physician well-being (November 2025), a university press release reporting a peer-reviewed randomised trial in NEJM AI, 2025 (DOI 10.1056/AIoa2501000). www.uclahealth.org/news/release/ucla-study-finds-ai-scribes-may-reduce-documentation-time
10. Khaleej Times. 'No side effects': Dubai authority lists words clinics cannot use in ads (September 2026). News report on Dubai Health Authority rules. www.khaleejtimes.com/uae/dubai-health-authority-claims-words-cannot-be-used-in-doctor-clinic-social-media-ads
11. Harvey. Vault: document review and due diligence. Company product page. www.harvey.ai/platform/vault
12. Anthropic. Harvey customer story. Company page. claude.com/customers/harvey
13. Anthropic. Claude legal plugin. Company product page. claude.com/marketplace/plugins/legal
14. WPP. WPP launches Agent Hub on WPP Open (2026). Company news release. www.wpp.com/en/news/2026/01/wpp-launches-agent-hub-on-wpp-open-providing-clients-with-access-to-advanced-agentic-ai

About Short Pencils

Short Pencils is an independent marketing and communications advisory for the Levant and the Gulf, led by Habib Rihana. We advise brands and institutions on marketing and communications strategy, briefs and research, and choosing and managing agencies. We also build AI tools that bring 30 years of experience in the region into everyday planning. Our proprietary AI tool for marketing and communications is in development and open for private testing.

shortpencils.com

Join our design partner programme

Our proprietary AI tool for marketing and communications is open for private testing. If you are interested in joining our design partner programme to test it, apply here.