MAY 21, 2026

Prompt engineering is dead. Work with AI as a senior partner instead.

The 2025 craft of writing tight, controlling prompts is now table stakes. Frontier models like Opus 4.7 reward a different mental model — and most teams haven't made the switch.

By **Dave Taylor**

Prompt engineering is over as a discrete skill. With Claude Opus 4.7 scoring 87.6% on SWE-bench Verified and delivering a 14% improvement in multi-step agentic reasoning while making a third of the tool errors of its predecessor, frontier models are no longer tools that need micromanaged instructions. They're senior collaborators that need context, edges, and a real thesis from you. Treat them like a junior assistant and you get junior output. Treat them like a senior partner and you get work that compounds.

Why prompt engineering became table stakes

The hiring-market evidence is already in. Microsoft's 2026 Work Trend Index, surveying 31,000 workers across 31 countries, ranks "prompt engineer" second-to-last among the new roles companies plan to add in the next 12-18 months. LinkedIn listings for the title have collapsed to near-zero in most markets, and the skill set has folded into broader AI-workflow and automation-design roles. That isn't because writing good prompts stopped mattering — it's because the bar moved. Knowing the difference between a chain-of-thought structure and a system-prompt template is now what we used to call "knowing how to format a Word document." Useful. Expected. Not a specialism.

The deeper shift is about who's doing the thinking. The 2025 mental model was *I describe the task in exhaustive detail, the AI executes it*. That worked when models were narrow and stateless. It doesn't work when the model can call tools, manage data, sustain a project across hours of work, and proactively verify its own output. Forcing senior-partner-grade software into junior-assistant patterns of interaction is the most common mistake we see in client engagements right now — and it's why a lot of AI pilots stall at "impressive demo" before they ship.

What changed: Opus 4.7 and the agentic shift

The April 2026 release of Claude Opus 4.7 is the clearest single marker of the transition. Anthropic built it explicitly for the work Opus 4.6 needed hand-holding through: long-running coding workflows, enterprise knowledge work across documents and spreadsheets, complex multi-tool operations. The model writes tests, runs them, and fixes the failures before declaring a task complete. It scores 77.3% on MCP Atlas and 78.0% on OSWorld — not the kind of capability you get out of a system that needs every step spelled out.

OpenAI's GPT-5.5 release earlier in the year did the same thing from a different angle, and Gemini 3.1 Pro followed. The pattern across all three labs is the same: the frontier is no longer "answer this question well", it's "carry this project across the next four hours without losing context". When the underlying capability changes that much, the way you interact with the system has to change too. Otherwise you're paying for senior-partner cost and getting senior-partner-restrained-by-junior-instructions output.

AI as a senior partner: the mental-model upgrade

The replacement mental model isn't complicated, but it's hard for people who built skill at the old game to drop. Interacting with a frontier model in 2026 should feel like running a problem set past a senior partner — someone with broad domain depth, a track record of

judgment calls, and the social standing to push back if you've got the thesis wrong. You don't micromanage that person. You give them context, the boundaries of the problem, your current thesis, and an invitation to argue with you.

The contrast matters because it changes what good input looks like. A 2025 prompt is closed-ended: *summarise this transcript into bullet points, no longer than 200 words each*. A 2026 question is open-ended: *I have a thesis that our marketing attribution is broken because most conversions come from sources the dashboard can't see. Read this attribution report, the last six campaign retros, and the customer-interview notes. Tell me whether my thesis holds up, what evidence points the other way, and what we'd need to test to be sure.* Different work. Different answer. Different value.

The Flashlight: conveying intent and edges

The first principle for working this way is what you might call the Flashlight. A flashlight beam has a clear centre and clear edges — bright in the middle, dark outside. Good senior-partner questions have the same shape. State your thesis upfront so the model has a centre to target. Then say what's out of scope so it doesn't drift into questions you weren't asking.

The opposite failures are equally bad. Open-ended questions with no centre ("what should I do about marketing?") produce vague summaries that name everything and recommend nothing. Closed-ended questions with no edges ("rewrite this in exactly five bullets") get you a competent five bullets and zero of the senior thinking you actually wanted. The art is the balance — intent that's clear enough to give the model a target, and edges that are firm enough to keep it inside the right problem space. The work-conversation in our heads sounds something like *"Here's where I'm pointing. Here's what I'm not asking. Now run."*

Complex question stacking

The second principle is harder to teach because it looks like the opposite of good prompting. The 2025 instinct was to break a complex task into clean, atomic steps and feed them one at a time. The 2026 instinct is the reverse: stack several hard, open-ended questions in a single request, and let the model synthesise an answer that holds all of them at once.

A worked example. If we're writing a PR FAQ for a software-and-hardware product launch, the old approach was "write a customer-facing FAQ, focus on benefits, use the brand voice". The new approach is something like *consider how the software experience feels to a customer who's never used our hardware; how the hardware constraints shape what the*

software can promise; how the launch needs to balance the emotional pull of the customer story against the internal logistics of who fulfils what when something breaks. Produce the FAQ that meets all three. The model is now reasoning, not retrieving. The output is shaped by tension, not by template.

Data, files, and implicit opinions

The third principle is about how you bring evidence into the conversation. Modern workspaces — Claude Code, Cursor, the various Anthropic and OpenAI workbenches — let you drop formal files (specs, code, spreadsheets) and informal ones (meeting transcripts, voice memos, Slack threads) into the same working folder. The mistake is assuming the model will read all of them with equal weight. It won't, unless you explicitly tell it to.

The pattern that works is: name the artefacts in your question, layer your soft opinions on top of the hard data, and give the model explicit permission to disagree with you. *Read the Q1 revenue review (formal), the three sales-team transcripts from last week (informal), and the proposed Q2 plan (draft). I'm leaning toward the bottom-up forecast, but I think the team's pessimistic. Pull together the cleanest thesis you can across all three sources. Push back on me where the evidence doesn't support what I'm saying.* That last line — the explicit invitation to argue — is what most prompts still don't include, and it's what unlocks the work you can't get from instruction-following.

What this means for Irish SMBs

For a 50-person Irish manufacturer, dental practice, or services firm trying to use AI in 2026, the decision isn't *which model do we adopt* — every frontier model is good enough now. The decision is *who in the team has internalised the senior-partner mental model*, because that's the bottleneck. McKinsey's 2025 State of AI survey shows the aggregate is encouraging: generative AI tools cut task time by 30-55% on content drafting, code generation, and data summarisation across the firms that have deployed them well. But the same survey flags the qualifier — those numbers cluster in organisations that already invested in governance, evals, and integration depth. They aren't picked up by buying the latest model.

In our work shipping AI agents for SMBs — lead generation, sales operations, custom internal agents — what separates a pilot that ships from a pilot that stalls usually isn't the technology. Our recent engagement with a 50-person recruitment and staffing firm lifted inbound calls 37% in an 8-week pilot. The pilots we see stall out aren't blocked by the model. They're blocked by whether the project leader brings junior-assistant mental models or senior-partner ones to every working session. The team that asks the agent to

execute the next ten steps gets ten plausible-looking steps and a stuck pilot. The team that asks the agent to *propose the three options it sees, flag the trade-offs, and tell us which one it'd push for and why* gets a working system in production. Same model, same data, different mental model. That's the whole game in 2026.

If you're trying to figure out where this mental-model shift lands for your own business, book a 30-min working session. We do this kind of scoping every week — the goal is to leave with one specific, smallest-useful AI slice that your team can ship in the next six weeks.

Related: this is the upstream framing for the AI Agent Reality Check series — Part 1: Treat your AI agent like a new hire covers the onboarding mechanics, and Part 2: The knowledge in your head is your AI bottleneck covers the tacit-knowledge problem behind it. And our case studies for what shipped AI work looks like in practice.