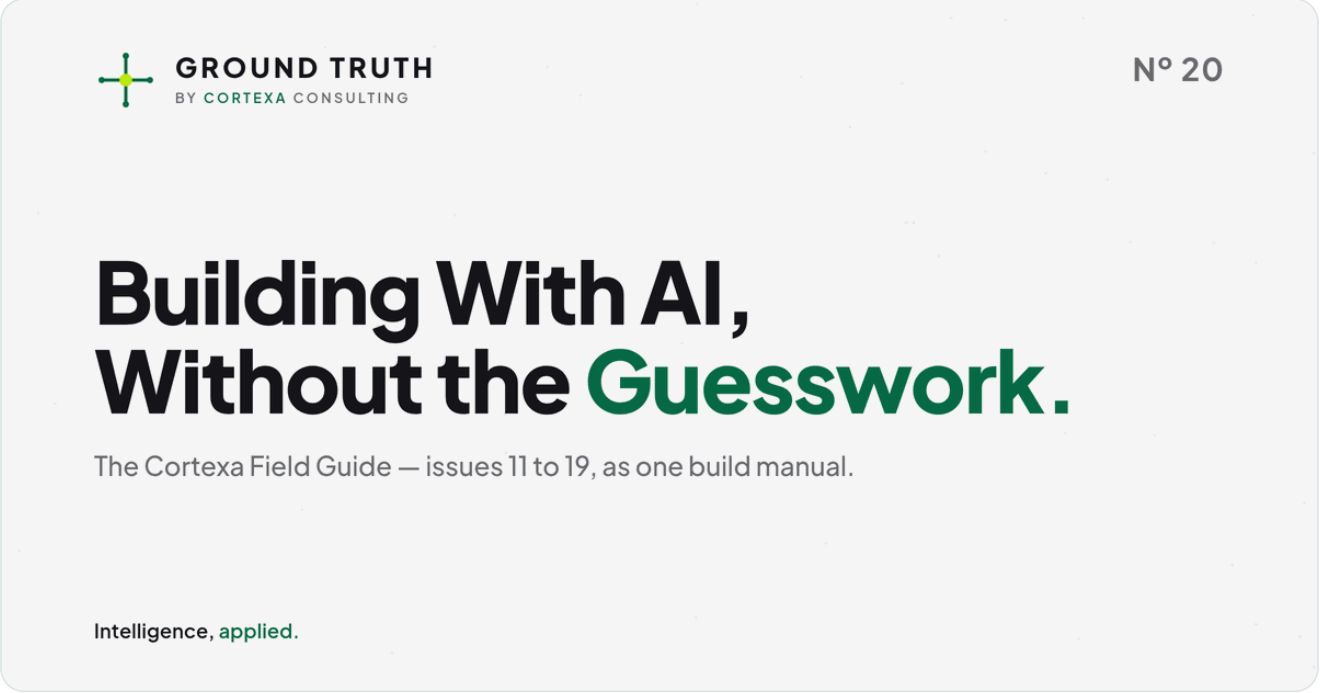
FIELD GUIDE

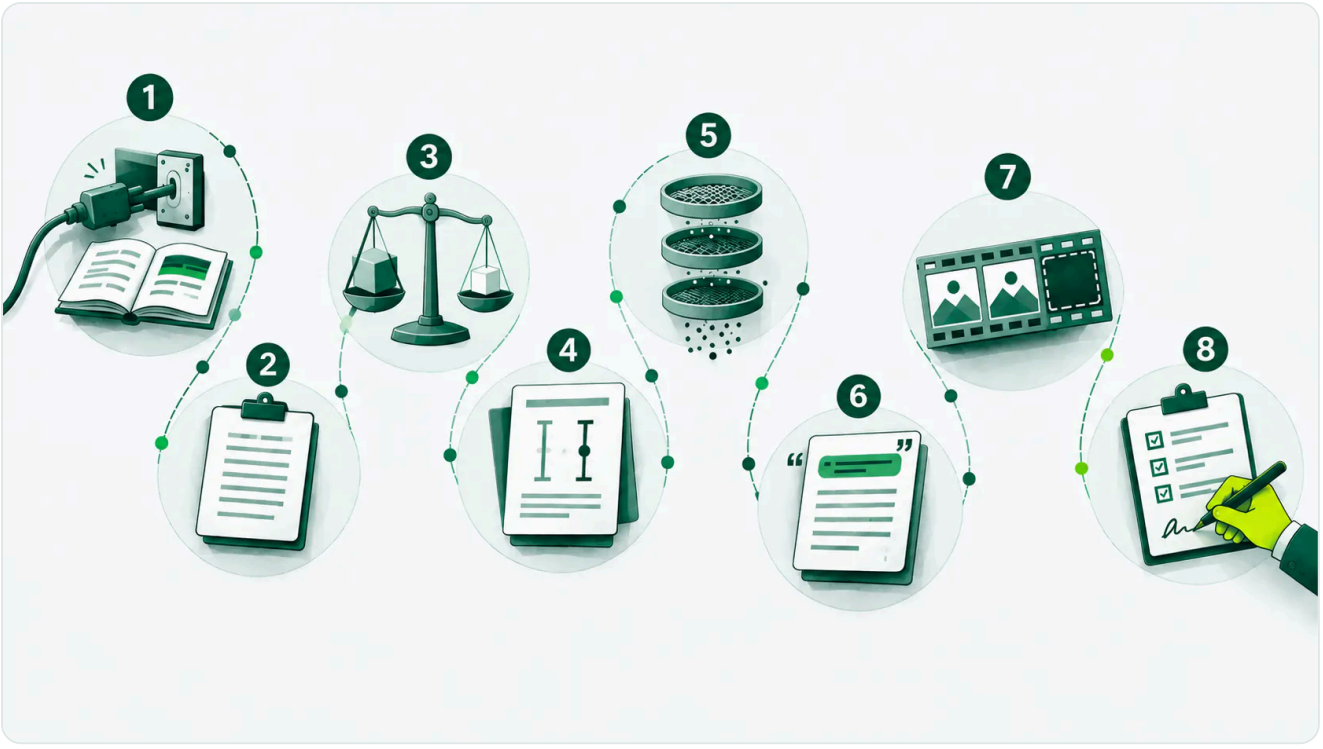
Building With AI, Without the Guesswork.

Nine explainers, one build manual: how AI reaches your tools, what the model choice costs, and what to check before a client's name goes on it.



*Intelligence, **applied.*** · September 15, 2026 · 20 min read · **www.cortexa-consulting.ai/insights/building-with-ai-without-guesswork**

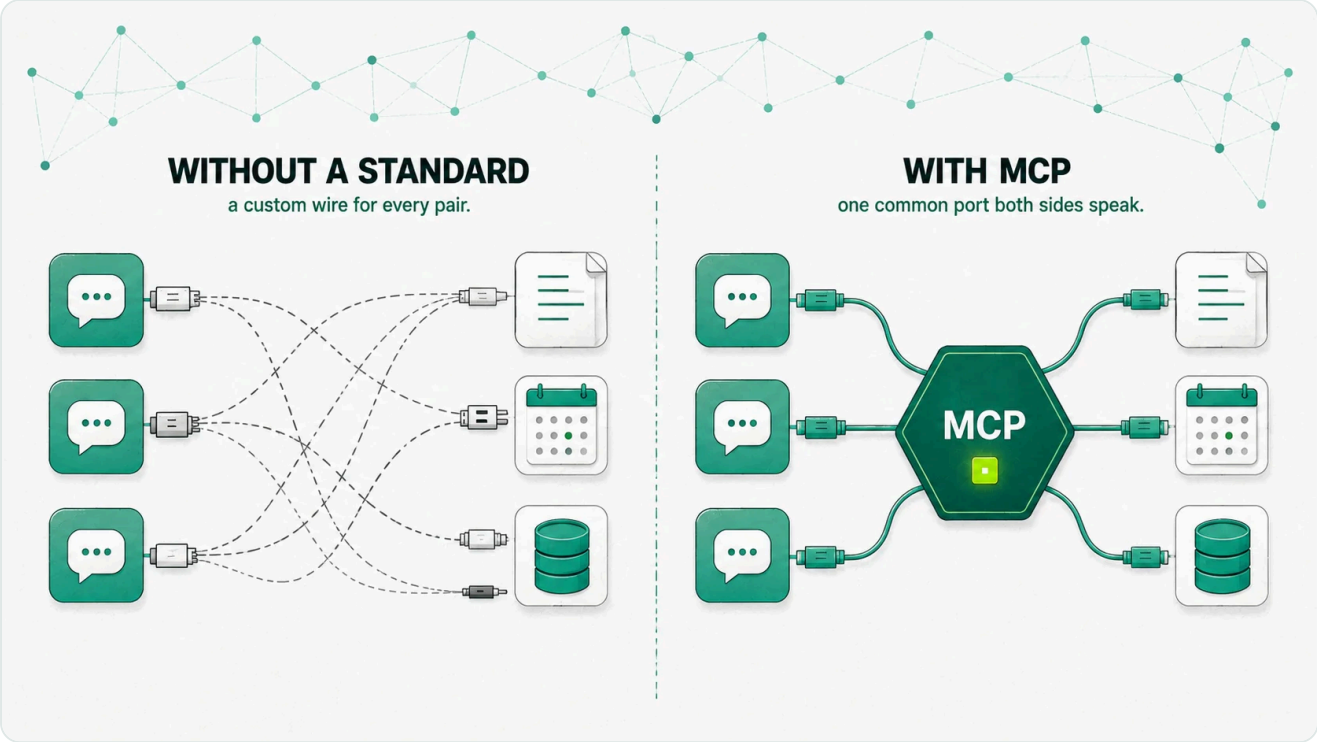
The first Field Guide, AI, Without the Hype[1], covered what Artificial Intelligence (AI) is and how it behaves. This one picks up where that left off and covers what it takes to build something with it. Nine explainers sit behind these eight chapters, and they arrived in the order the work itself arrives: first you wire the thing into your tools, then you give it a job, then you argue about which model, and somewhere near the end someone asks whether it is safe to put a client's name on the output. Read it as a build manual. Each chapter names the decision, shows what the evidence supports, says what it means when you are scoping and quoting, and points back at the explainer it came from when you need the detail.



The eight chapters as one run. They are not options to choose among; you walk them in order, and a person signs the end of it.

1. Wiring it up

A model on its own knows nothing about your client. Two wires change that. The first reaches your tools[3]: the Model Context Protocol (MCP) is an open standard for connecting an assistant to software, and Anthropic describes it as a universal, open standard for connecting AI systems with data sources[12]. OpenAI supports it too, which is what turned it from one vendor's idea into a port worth building against. The second wire reaches your knowledge[4]: text becomes coordinates, and the system retrieves the passages nearest the question. Both wires do the same job. They put the right context in front of the model at the moment it answers.



One standard port instead of a custom wire per app. That is the whole idea behind MCP.

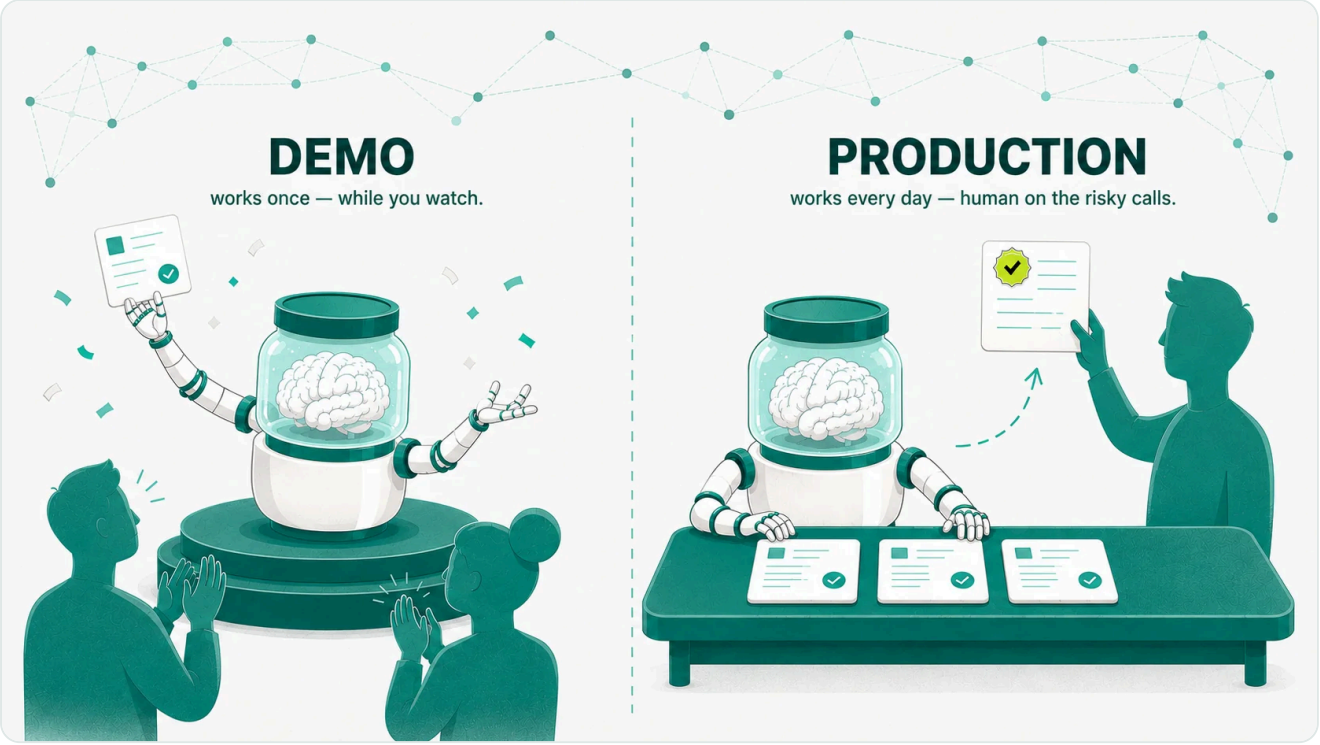
- Retrieval is measurably better than keyword matching for this. A purpose-built dense retriever beat a strong keyword baseline by 9% to 19% absolute on top-20 passage retrieval accuracy[13]. That gap is why "search our documents" became a retrieval project rather than a search-box project.
- A protocol is a connector, and it carries no judgment about what should be connected. The same port that reaches a calendar reaches a production database. OWASP files that risk as excessive agency[15], and the fix is scoping the permissions rather than trusting the model to be careful.

> **CORTEXA'S READ**
>
> When a client says "train the AI on our data," they almost always mean this chapter rather than training. Wiring is cheaper, updates the moment a document changes, and every answer can point back at the passage it came from.

2. Giving it a job

An agent is a model handed a goal and the means to act on it. The demos are spectacular and the failure mode is boring: a system that works in a scripted run and wanders on the real thing. Anthropic's own guidance is unfashionably plain. The teams that succeed find the simplest solution possible, and only increase complexity when needed[14], and many jobs

never need the full agent loop. A shipped agent[2] usually turns out to be a narrow task, a small set of tools, and a person who signs the result.



The difference between a demo and a shipped agent is rarely intelligence. It is scope, and who signs.

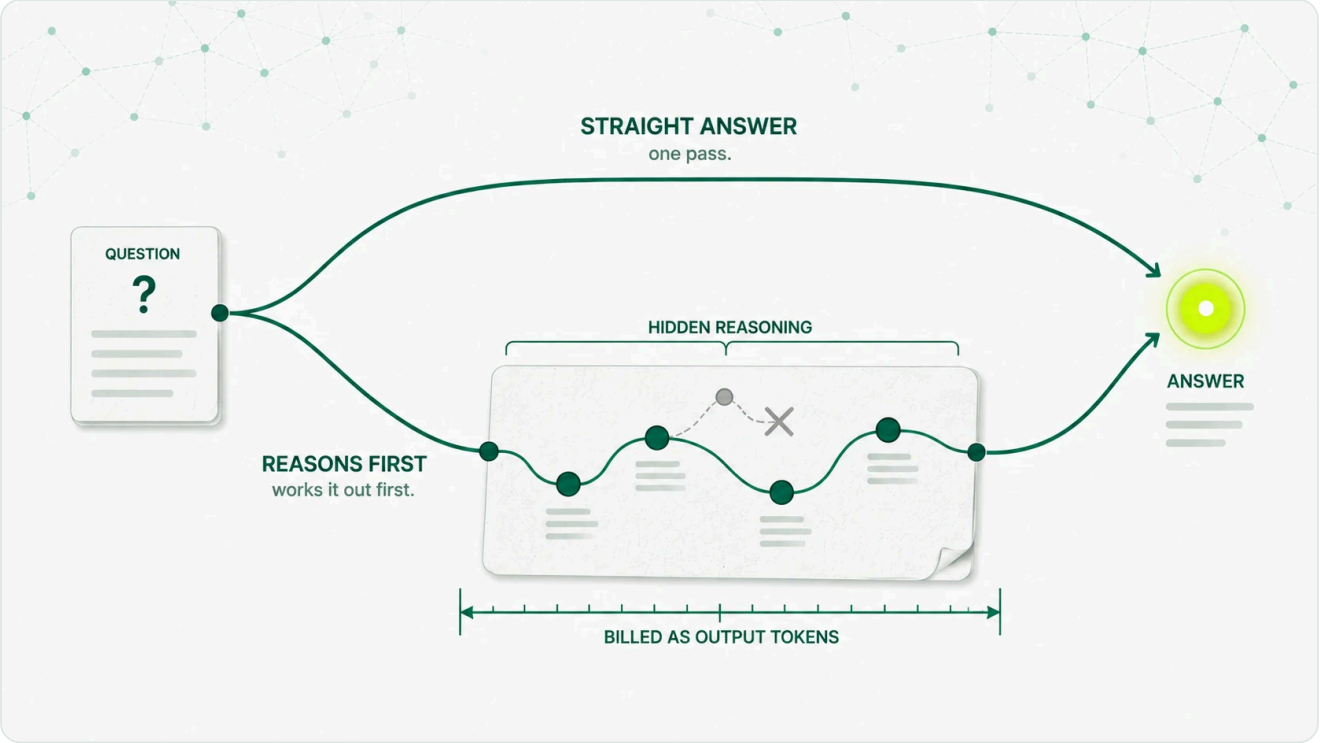
- Scope the permissions before the capability. OWASP's excessive agency[15] entry covers exactly this: too much functionality, too many permissions, too much autonomy. Every tool you connect is a thing the agent can do at three in the morning without asking.
- Governance is a control requirement. The NIST AI Risk Management Framework[16] treats measurement and oversight as part of the system rather than paperwork bolted on afterward. Write the oversight into the statement of work, not into the retrospective.
- Decide what "done" means before you build. Without a test you agree on up front, "the agent works" becomes an argument about vibes, which is the argument evals[11] exist to settle.

CORTEXA'S READ

Ask a vendor which steps a human still signs. If the answer is none, you are looking at a demo. If the answer is specific, you are looking at something that has met production.

3. Choosing the model

Two settings drive most of an AI feature's bill, and neither is the sticker price per token. The first is whether you reach for a reasoning model[6], which works the problem out privately before it answers. That hidden work is billed: OpenAI is explicit that reasoning tokens are billed as output tokens[17] even though you never see them. The second is whether you reuse your standing context. Prompt caching[5] reads the unchanging part of a prompt once and resumes from it, and Anthropic prices a cache read at roughly a tenth of the base input rate.



The thinking is real work, and you pay for it whether or not the question needed it.

- Reasoning is a dial, and it can be turned up too far. Researchers have documented models spending far more computation than easy problems warrant, calling it overthinking[18]. Pointed at a simple lookup, a reasoning model costs more and takes longer for an answer that is no better.
- The private chain is not a confession. Anthropic found that reasoning models don't always say what they think[19], so a visible chain of steps is a useful artifact rather than an audit trail. Do not sell it to a client as one.
- Caching pays when a large standing prompt gets reused inside a short window. Structure the prompt with the stable material first, and the discount is available. Retrofit it later and you are rewriting the feature.

CORTEXA'S READ

Quote AI features in tokens and behavior, not in a monthly guess. The two questions that move the number most are whether the model thinks before answering and whether the standing context gets reused.

4. Reading the claims

Every vendor arrives with a chart. A benchmark[9] is a fixed exam, and the score says how a model did on that exam under one set of choices. Read it the way you would read a single test score for a candidate: narrow, and easy to over-interpret. Stanford's HELM makes the honest version of this visible by reporting accuracy, robustness, fairness, efficiency[20] and more alongside each other, rather than collapsing a model into one number.

A high score is a strong general result. An excellent key still won't turn a lock cut to a different shape.

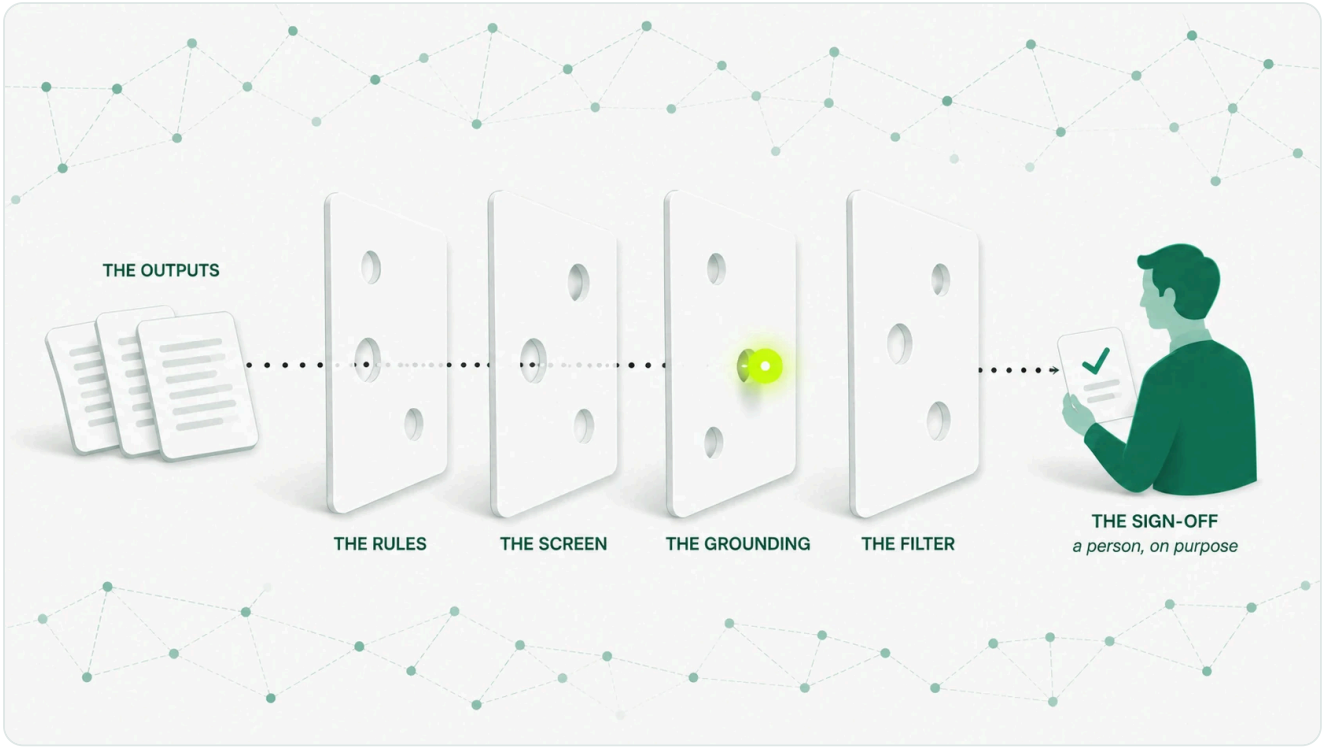
- The gap at the top is often inside the noise. Anthropic's own evaluation work argues for adding error bars to evals[21] and treating scores as statistical estimates. Draw the interval every score deserves and a one-point lead can vanish.
- The exam can leak into the study material. A survey of benchmark data contamination[22] documents test items turning up in training data, which inflates a score without improving the model.

- Leaderboards have their own dynamics. The Leaderboard Illusion[23] paper describes how selective reporting and unequal access distort arena rankings. Treat a rank as a shortlist, then run your own small test on the client's real task.

> **CORTEXA'S READ**
>
> The score that should decide a client's stack is the one you produced on the client's own work. Everything upstream of that is a shortlist.

5. Safe to publish

Asked whether an AI feature can be guaranteed never to say anything off-brand, the honest answer is no, and the useful answer is layers[8]. Give it the tone and the refusals up front. Screen what comes in, ahead of the model. Hand it approved material to quote from. Scan the reply on the way out. Then put a person in front of anything carrying a client's name. Each control catches a share of what reaches it. Stacked, with different gaps, very little threads all of them.



One screen lets the odd thing through. Four screens with the holes in different places stop almost everything.

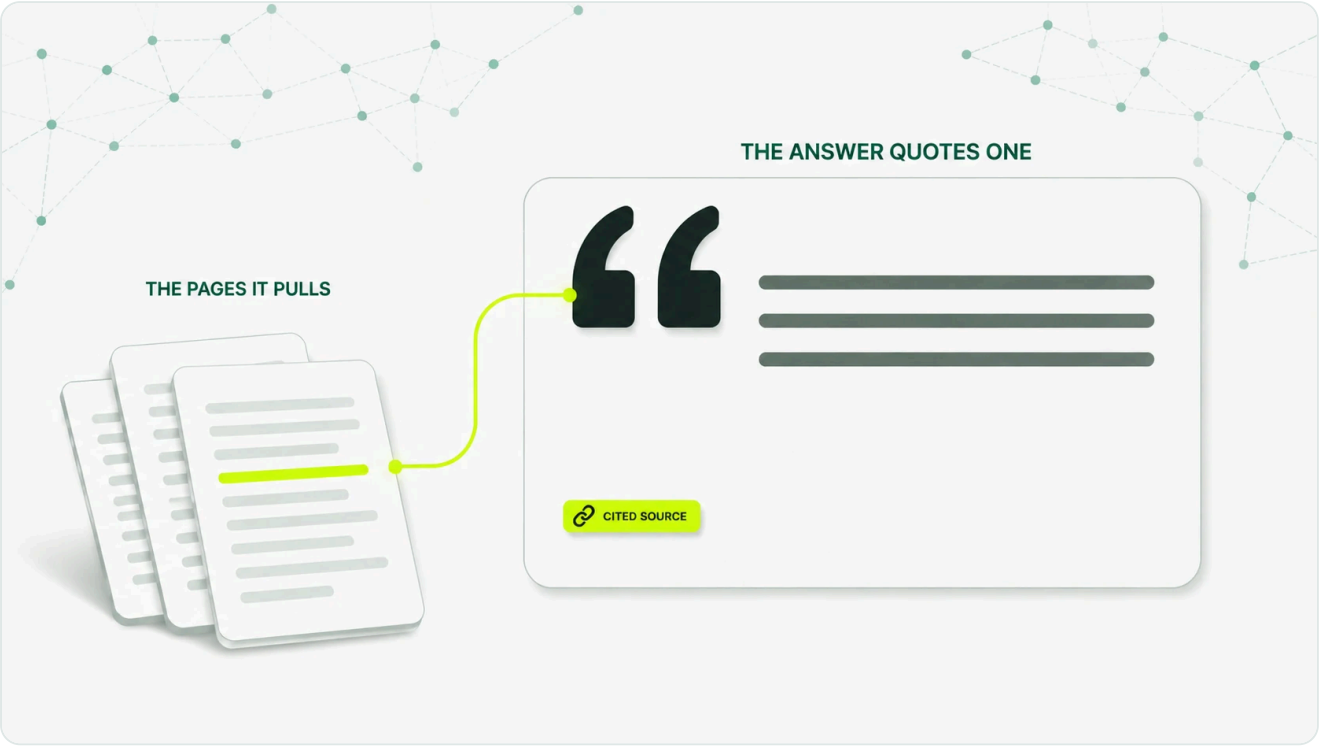
- Prompt injection is the standing risk and it has no complete fix. OWASP ranks prompt injection[24] first among risks to applications built on language models, which is why input screening and output filtering are separate layers rather than one setting.

- A filter catches categories, and your brand is not a category. Toxicity classifiers do not know that a client never says “cheap.” That gap is what the human on the last mile is for.

> **CORTEXA'S READ**
>
> Sell “reviewed and reduced,” never “impossible.” A stack of controls with a named person on the last mile is a promise you can keep, and it survives the one bad output that any honest system will eventually produce.

6. Being found

Clients have started asking how they show up when someone puts the question to an assistant instead of a search box. The mechanism is less mysterious than the pitches suggest. Assistants answer by retrieving live pages and quoting them, and Google says its AI features rely on our core Search ranking systems to retrieve relevant, up-to-date web pages[25]. So the route to being cited[7] runs through being retrievable first and quotable second.



Many pages get pulled. The answer quotes the clearest one, and names it.

- Credibility moves are the ones that measure. The paper that gave Generative Engine Optimization (GEO) its name measured nine tactics against live answers. Sourcing, quotation and hard numbers came out on top, lifting a page's visibility by up to 40%[26] in generative answers.

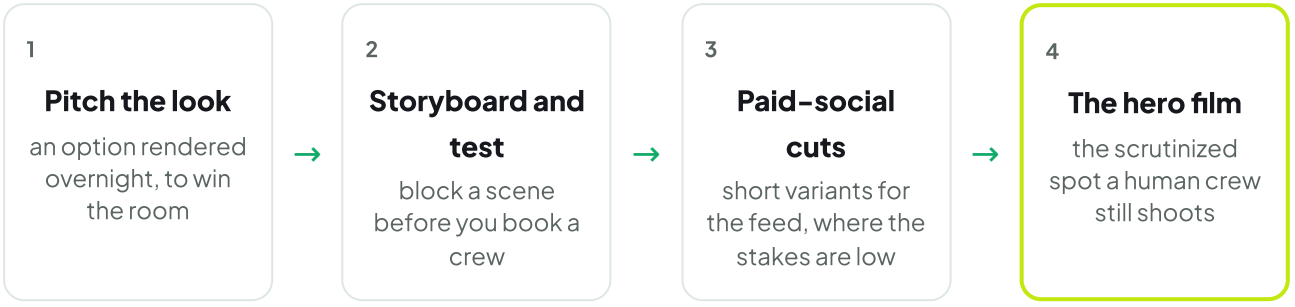
- What worked on a search engine does not transfer. The same paper found keyword stuffing[26] gave little to no improvement. When a vendor pitches a secret setting that unlocks the assistants, the pitch is the product.

> **CORTEXA'S READ**
>
> Being quotable and being trustworthy have converged into the same job. Answer the question in the first line, and name the primary you got the number from.

7. Making the work

Generative video arrived fast enough to reset the question every quarter. Google's Veo can generate cinematic video with audio[27], which makes it genuinely useful early in the process. It stays risky at the finish, because holding a subject steady across a longer take is the open research problem: the survey literature notes video generation demands not only high-quality individual frames but also strong temporal coherence[28]. What holds up in practice is generative video for the storyboard[10] and a human crew for the film that ships.

WHERE GENERATIVE VIDEO EARNS ITS PLACE



The first three stages are ready today. The last one is where the tells still show.

- The tells cluster where an audience lingers. Coverage of Coca-Cola's all-AI holiday spot cataloged an uncanny gloss, diffuse lighting, rubbery facial features, frequent cuts, and a lack of spoken dialogue[29]. Those are the artifacts of a long take rather than a short one.
- The honest nuance cuts both ways. In the same coverage, viewers who were not told the ad was AI-made rated the ad similarly to the original 1995 spot[29]. The reputational risk lives closer to the disclosure than to the pixels.

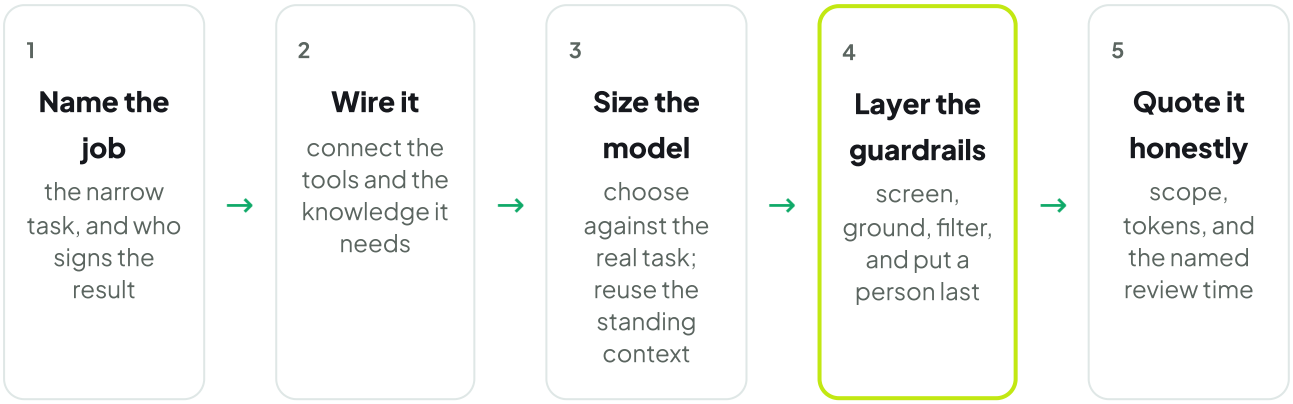
CORTEXA'S READ

Point generative video at the work that gets thrown away: the option you show to win the room, the cut you test and discard. Keep a crew on the thing that carries the logo.

8. The agency playbook

Seven chapters of decisions collapse into one working order. Establish what the client wants the system to do, and whether a person signs the result. Wire it to the tools and the documents it needs. Choose the model against the job rather than the leaderboard, and structure the prompt so the standing context is reused. Put the layers in before anything is published, and price the human review as a line item rather than absorbing it. The sequence matters more than any single choice in it, because each step constrains the next.

THE ORDER THE DECISIONS COME IN



Change the job and you walk the path again; every later step depends on the first.

- Scope the narrow task, not the category. “An agent for the account” is unquotable. “Draft the weekly performance summary from these three reports, reviewed before it sends” is a piece of work with an edge you can price.
- Put the review in the estimate. The automated layers cost little; the one that catches what slips past them is somebody's booked hours. Budgeting for it is what makes the rest of the promise honest.
- Bring a test, not an opinion. Run the client's real task against two candidates and show the result. It ends the leaderboard argument in a meeting rather than a memo.
- Say what you do not know. The disclosure that ages best is the one that named its limits while the work was being scoped, which is the whole reason the first Field Guide[1] exists.

CORTEXA'S READ

None of this needs a research team. It needs a straight answer from the client about the job, and the discipline to walk the path in order. That is the whole of the guesswork removed.

Sources

1. Cortexa Ground Truth Nº 10 — “AI, Without the Hype — The Cortexa Field Guide”
   https://www.cortexa-consulting.ai/insights/whats-real-in-ai
2. Cortexa Ground Truth Nº 11 — “Which AI agents actually ship?”
   https://www.cortexa-consulting.ai/insights/agents-that-actually-ship
3. Cortexa Ground Truth Nº 12 — “MCP is just a universal plug”
   https://www.cortexa-consulting.ai/insights/what-is-mcp
4. Cortexa Ground Truth Nº 13 — “Vector search is just a map of meaning”
   https://www.cortexa-consulting.ai/insights/vector-databases-and-embeddings
5. Cortexa Ground Truth Nº 14 — “Prompt caching is just a bookmark”
   https://www.cortexa-consulting.ai/insights/what-is-prompt-caching
6. Cortexa Ground Truth Nº 15 — “Reasoning models show their work”
   https://www.cortexa-consulting.ai/insights/when-reasoning-models-are-worth-it
7. Cortexa Ground Truth Nº 16 — “To be cited by AI, be worth quoting”
   https://www.cortexa-consulting.ai/insights/getting-cited-by-ai-answers
8. Cortexa Ground Truth Nº 17 — “Guardrails work in layers”
   https://www.cortexa-consulting.ai/insights/keeping-ai-outputs-brand-safe
9. Cortexa Ground Truth Nº 18 — “Read an AI benchmark like a test score”
   https://www.cortexa-consulting.ai/insights/how-to-read-an-ai-benchmark
10. Cortexa Ground Truth Nº 19 — “AI video is ready for the storyboard, not the shoot”
    https://www.cortexa-consulting.ai/insights/ai-video-for-ads
11. Cortexa Ground Truth Nº 8 — “Evals, in plain English”
    https://www.cortexa-consulting.ai/insights/evals-in-plain-english
12. Anthropic — Introducing the Model Context Protocol
    https://www.anthropic.com/news/model-context-protocol
13. Karpukhin et al. — Dense Passage Retrieval for Open-Domain Question Answering (preprint)
    https://arxiv.org/abs/2004.04906
14. Anthropic — Building effective agents
    https://www.anthropic.com/research/building-effective-agents
15. OWASP — LLM06:2025 Excessive Agency (Top 10 for LLM Applications)
    https://genai.owasp.org/llmrisk/llm062025-excessive-agency/
16. NIST — AI Risk Management Framework
    https://www.nist.gov/itl/ai-risk-management-framework
17. OpenAI — Reasoning models (API guide)
    https://developers.openai.com/api/docs/guides/reasoning
18. Chen et al. — Do NOT Think That Much for 2+3? On the Overthinking of o1-Like LLMs (preprint)
    https://arxiv.org/abs/2412.21187
19. Anthropic — Reasoning models don't always say what they think
    https://www.anthropic.com/research/reasoning-models-dont-say-think

20. Stanford CRFM — HELM (Holistic Evaluation of Language Models)
https://crfm-helm.readthedocs.io/en/latest/

21. Miller (Anthropic) — Adding Error Bars to Evals (preprint)
https://arxiv.org/abs/2411.00640

22. Xu et al. — Benchmark Data Contamination of Large Language Models: A Survey (preprint)
https://arxiv.org/abs/2406.04244

23. Singh et al. — The Leaderboard Illusion (preprint)
https://arxiv.org/abs/2504.20879

24. OWASP — LLM01:2025 Prompt Injection (Gen AI Security Project)
https://genai.owasp.org/llmrisk/llm01-prompt-injection/

25. Google Search Central — Optimizing for generative AI features on Google Search
https://developers.google.com/search/docs/fundamentals/ai-optimization-guide

26. Aggarwal et al. — GEO: Generative Engine Optimization (KDD 2024; preprint)
https://arxiv.org/abs/2311.09735

27. Google DeepMind — Veo (model overview)
https://deepmind.google/models/veo/

28. A Survey: Spatiotemporal Consistency in Video Generation (preprint)
https://arxiv.org/abs/2502.17863

29. eMarketer — Coca-Cola used AI to generate its holiday ad campaign
https://www.emarketer.com/content/coca-cola-used-ai-generate-its-holiday-ad-campaign