FIELD GUIDE

AI, Without the Hype — The Cortexa Field Guide.

Nine explainers, one decision tool: when a client asks about AI, here's what's real, what to scope, and the honest version — chapter by chapter.

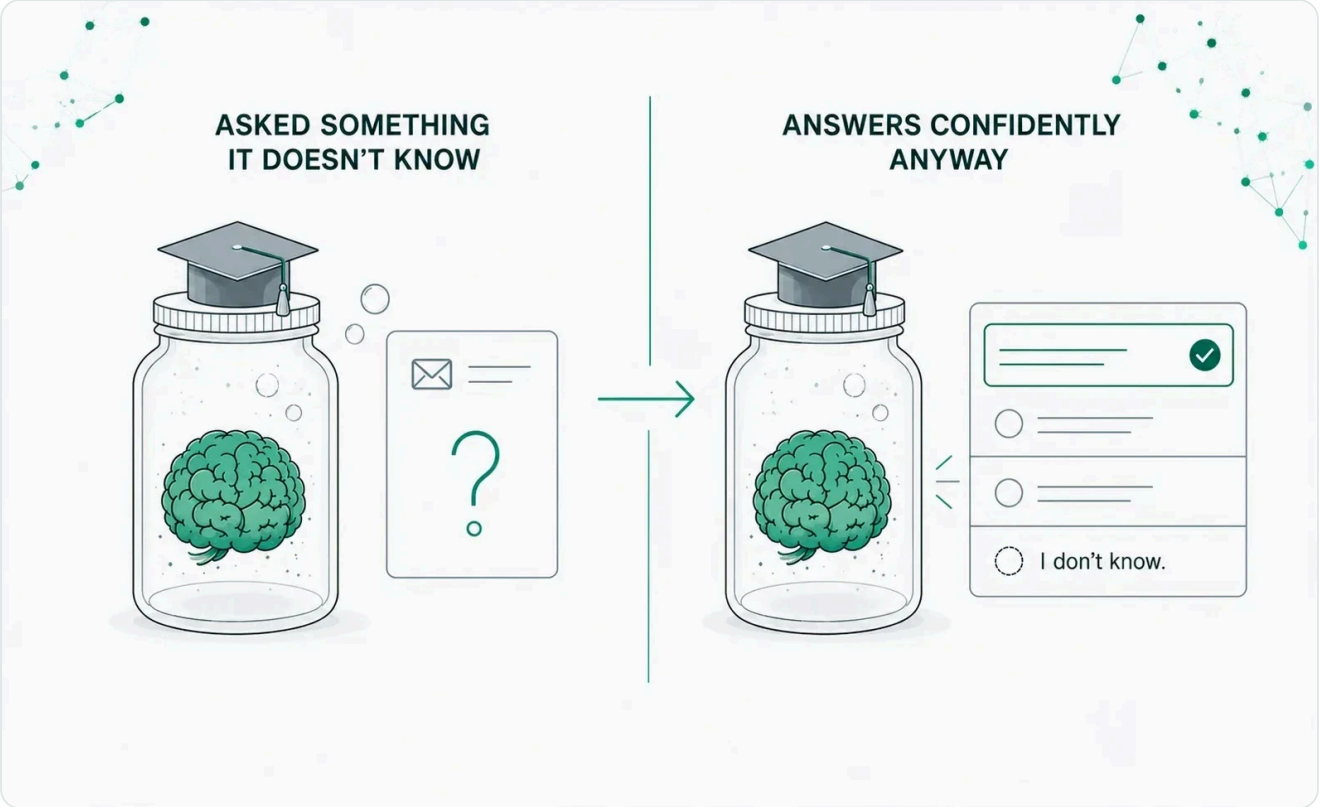
Intelligence, ***applied.*** · July 21, 2026 · 18 min read · **www.cortexa-consulting.ai/insights/whats-real-in-ai**

Every Ground Truth so far has answered one question. This is the map that connects them. Moving forward, this will be the cadence: for every ten Ground Truths, we will ship a Field Guide like this one — tying together the previous 8–9 explainers, and, over time, threading in earlier Field Guides too. When a client asks about Artificial Intelligence (AI) — “can it look at our slide deck?”, “should we fine-tune?”, “is the agent going to run the account?” — you don't need a research team; you need a straight answer from the client and an idea of how to scope the challenge. That's this guide: the nine explainers, synthesized into one decision tool. Every chapter is the same shape — it describes the topic/concept, explains what's real, and what it means when you're quoting the work — with a link down to the full explainer when you want to go deeper.

The whole guide in one view: nine explainers, one connected map. Everything orbits the brain in a jar — a confident guesser — the fact the rest of the guide is built to manage.

1. How AI actually works

Strip away the mystique and a Large Language Model (LLM) does one thing: it predicts the next likely chunk of text, over and over. Most of the time the likeliest words are also true — which is why it's so useful — but when it's unsure it doesn't stop or flag it. It produces a fluent, confident guess anyway. That isn't a bug to patch; it's how the machine works.



Asked something it doesn't know, the model answers confidently anyway.

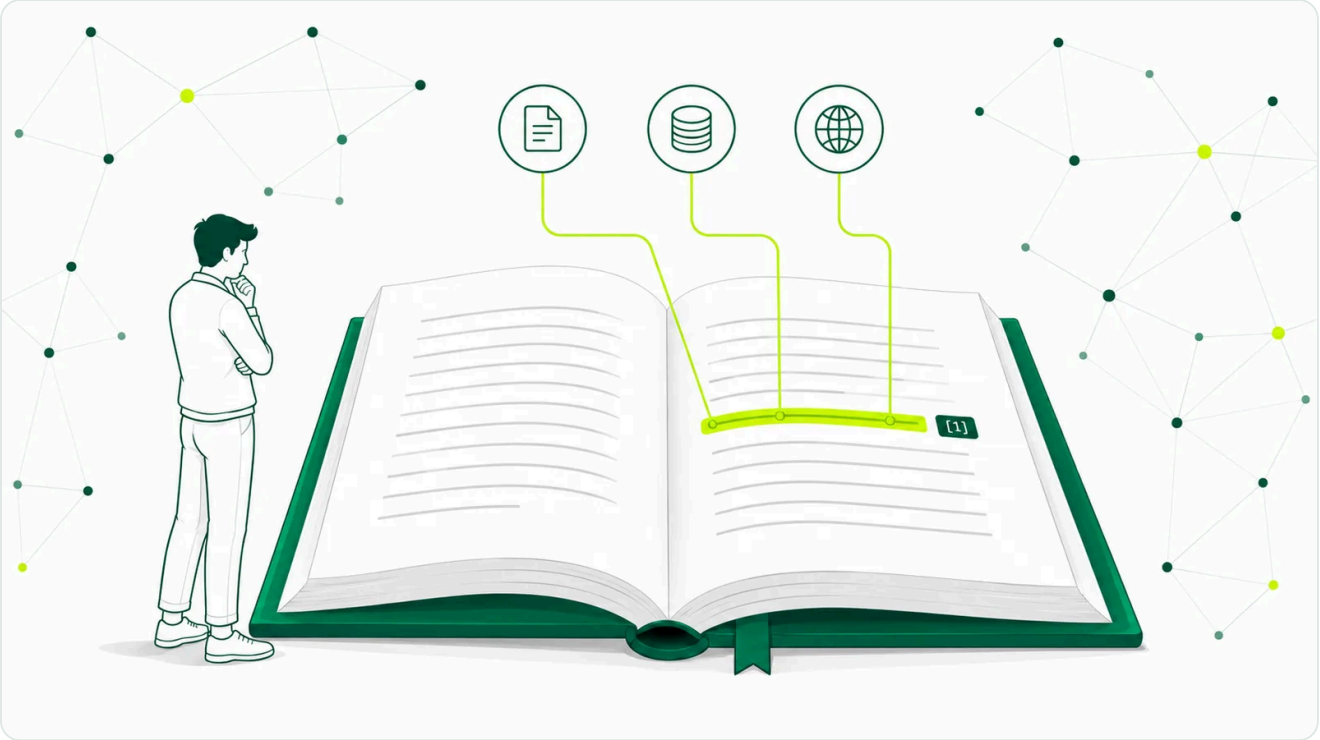
- It's **predicting, not remembering**. A major research review of hallucination with AI traces the failure to this[1] — the model reconstructs a plausible answer rather than reading from a source of truth.
- A confident tone tells you nothing about accuracy. There's no little voice inside saying "I don't actually know this."
- **Tokens** are the currency in the game of AI/LLMs. A token — roughly ¾ of a word — is what the context window (the model's short-term memory) is measured in, and what shows up on the invoice[2].

> **CORTEXA'S READ**
>
> The single most useful thing to know about AI is that it's a confident guesser, not a database. Every other chapter in this guide is a way to manage that one fact.

The "so what" for agencies: when a client says "the AI got the numbers wrong," the honest reframe is that it guessed confidently[3] because nothing grounded it — the fix is sources plus a review step, not a better prompt. And when you scope an AI feature, scope it in tokens[4]: input size × output length × how often it runs is the number that drives the bill.

2. Giving it knowledge

The most common client ask — “can we train it on our data?” — and it almost never requires retraining the model. It means: make it answer from our material, not from whatever it half-remembers. Three tools get conflated: **prompting** (tell it clearly what you want), **retrieval** (hand it the right documents to answer from — RAG), and **fine-tuning** (actually retrain it). They're a ladder — and you climb only when the rung below isn't enough to answer your requirements.



RAG turns the closed-book guesser into an open-book student — answering from the source.

- Grounding is the biggest, cheapest accuracy win. The foundational retrieval research[5] showed a model that pulls from an external source beats leaning on memory — especially for facts that change.
- Nine out of ten “custom AI” asks are RAG, not fine-tuning[6]. Reach for fine-tuning only when you need a consistent voice or output shape at scale — it changes style, not facts[7].

> **CORTEXA'S READ**
>
> “Train it on our data” almost always means “make it answer from our data.” That's grounding, not training — faster to stand up, cheaper to run, and updatable by changing the documents, not the model.

The “so what” for agencies: when a client says “fine-tune it,” translate it into a grounding project first — which documents should it answer from? You'll scope something cheaper,

faster, and updatable, and sound like the person who actually knows how this works.

3. Talking to it

"Should we hire a prompt engineer?" is usually the wrong question. The teams getting good work out of AI aren't chanting magic words; they're writing a clear brief. A good prompt **is** a clear brief: be specific about what "good" looks like, give the context a newcomer would need, and show a couple of examples of the output you want.

Same request, two ways to ask. The clear brief is the whole "technique."

- Clarity beats cleverness. The model-builders' own rule: show your prompt to a colleague with minimal context — if they'd be confused, the model will be too[8].
- Examples are the most reliable lever — a few well-chosen examples steer output more than any clever phrasing you can't repeat. It's a clear brief, not a spell[9].

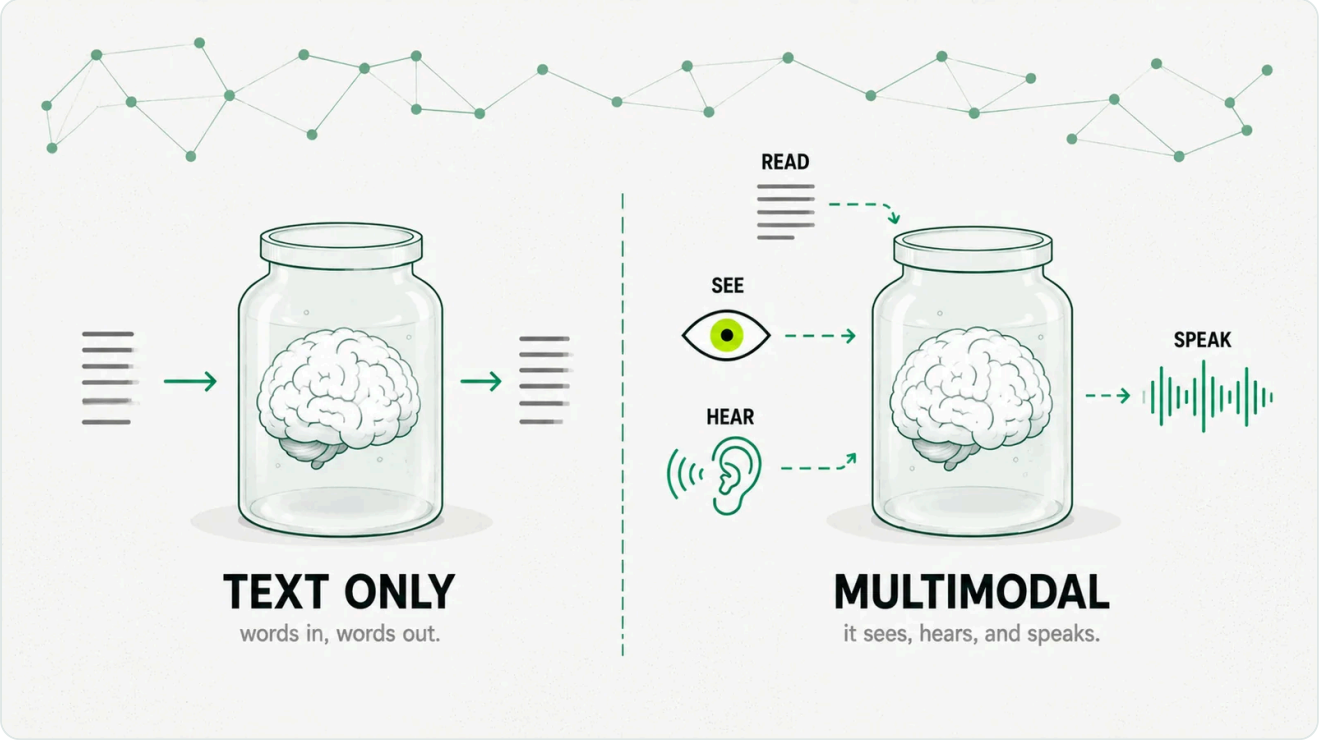
> **CORTEXA'S READ**
>
> "Prompt engineering" isn't a priesthood — it's the skill your best leaders and tech writers already have: saying precisely what you want, with the context and examples that make it unambiguous.

The "so what" for agencies: "we need a prompt engineer" is usually a coaching problem, not a hiring gap. Build a small library of on-brand example outputs — it's reusable, teachable,

and improves every prompt that uses it.

4. Giving it senses

When a client asks the AI to “look at” a layout or “listen to” a call, that's a **multimodal** ask — the model working in more than text. The quiet wins are the useful ones: read the screenshot, check a layout against the brief, pull the key points from a recording. The loud demos — the instant hero image — rarely ship as-is.



Multimodal is the same brain given eyes and ears — and a voice.

- The commercial frontier models really are multi-sense: OpenAI describes GPT-4o (o = Omni) as one model that accepts any combination of text, audio, image, and video[10].
- Always ask which sense, and which direction[11]. “Can it read our decks?” (understanding) is a different, easier project than “can it make our decks?” (generation) — and a different model may win at each.

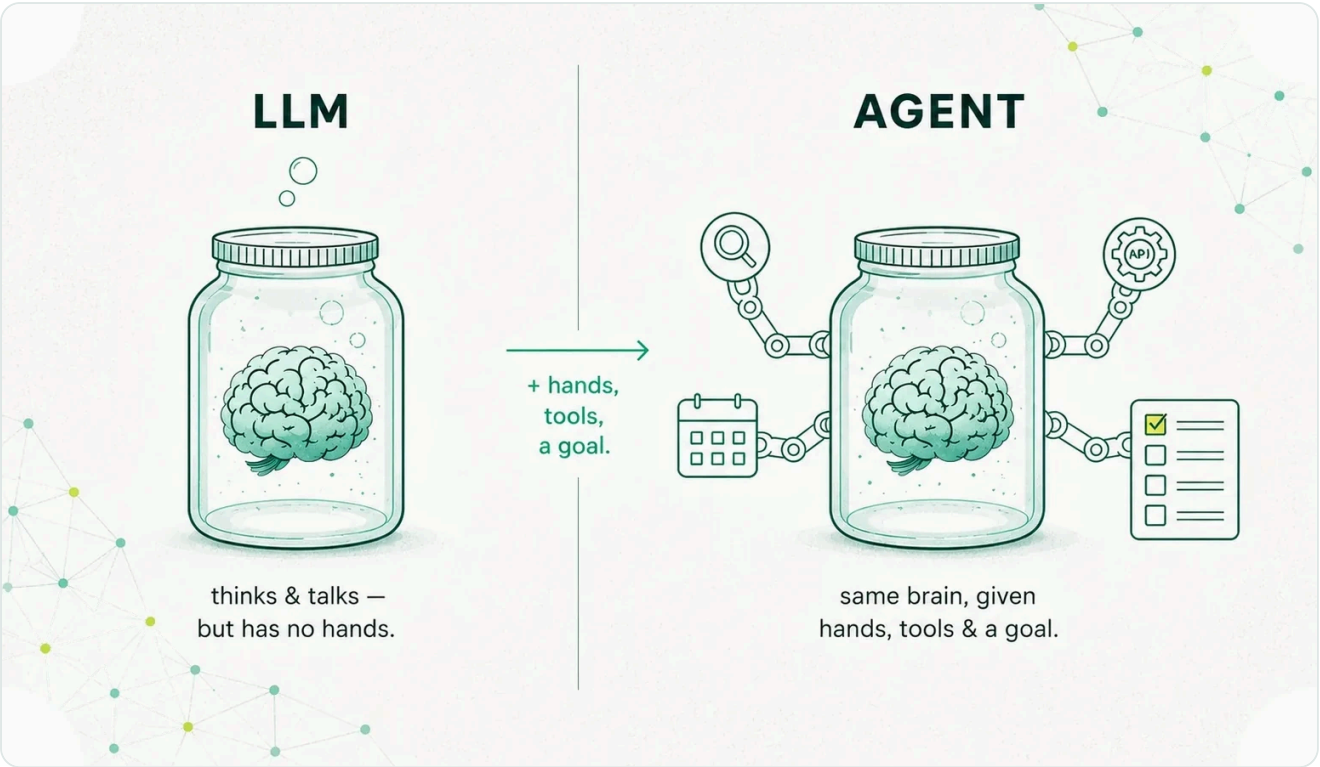
> **CORTEXA'S READ**
>
> The party trick is a hero image nobody can ship as-is; the real value is a model that finally sees what you're pointing at. Scope the seeing, not the spectacle.

The “so what” for agencies: multimodal asks are usually the cheap, useful kind — describe this image, summarize this recording, check these layouts against the brand guide. Scope

those first; they ship. Be precise about direction in the quote.

5. Giving it hands

An **agent** is that same guessing brain given hands, tools, and a checklist — it can take multi-step actions toward a goal, not just answer. That's the real leap. But it's a capable new intern, not an autopilot: fast, tireless, and occasionally confidently wrong.



An LLM thinks and talks. An agent is the same model given hands, tools, and a goal.

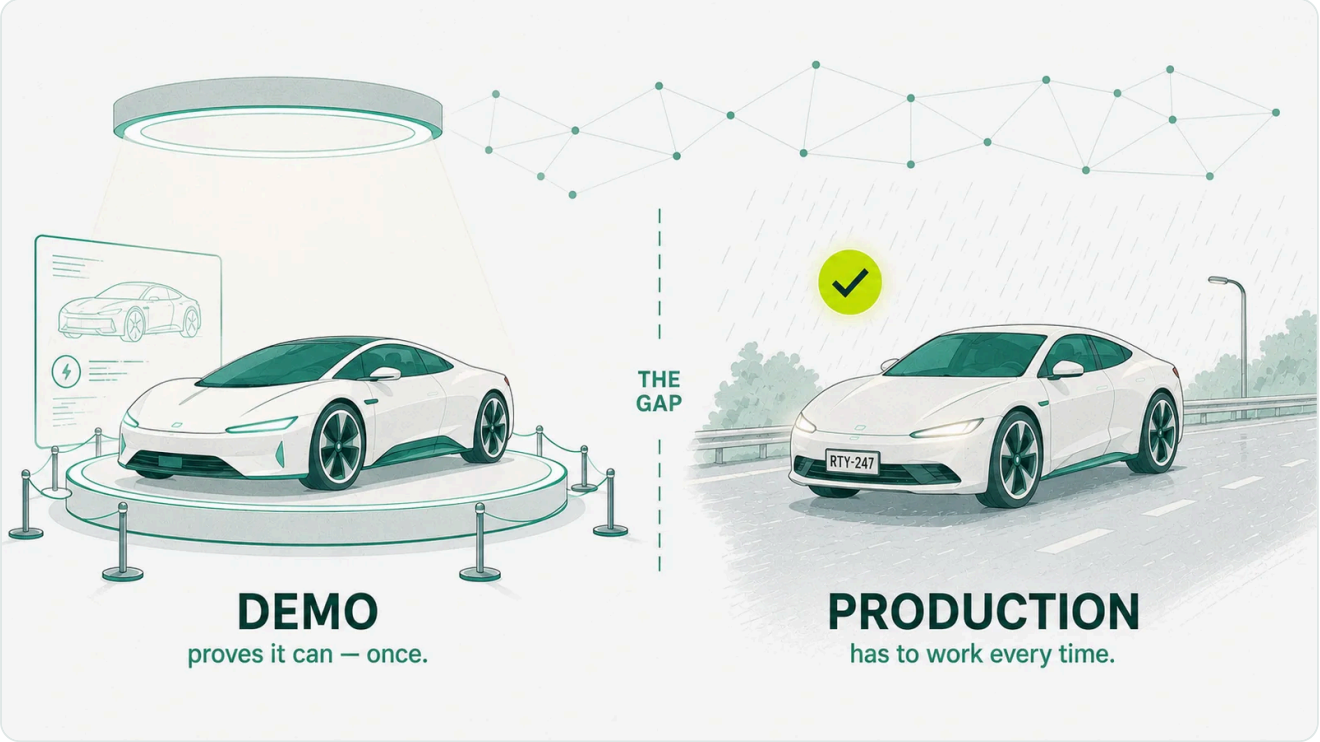
- The patterns that actually ship are simple and orchestrated — Anthropic's guidance argues for the simplest thing that works[12] over elaborate autonomy. It's an intern, not an autopilot[13].
- For anything with stakes, a human stays in the loop by design — governance like NIST's AI Risk Management Framework[14] treats human oversight as a control requirement, not as something optional.

> **CORTEXA'S READ**
>
> The useful question isn't "can we build an agent?" — it's "which repetitive, checkable step would we hand a sharp new intern, with someone reviewing the output?" Start there.

The “so what” for agencies: when a client asks for “an AI agent,” translate it into a specific, checkable task before quoting scope — most “agent” asks are one or two automatable steps with a human checking the result.

6. Shipping it

“We saw it work” greenlights the project — and quietly sets it up to stall out or fail. A demo proves the tool **can** do something once, on a good day, with a set stage. Production means it **does** the thing every time, even with messy inputs, and without someone standing by. That gap is where most AI projects live — and it's the most time-consuming, and drives the most cost/billable hours.



A concept car dazzles on a turntable. A production car has to start every morning — in the rain.

- You can't ship what you can't measure. The guidance is blunt: define success criteria, then build evaluations against them[15]. “It looked right in the demo” isn't a metric; a score on a real test set[16] is.
- Two levers do most of the reliability work: grounding (chapter 2) and guardrails — input/output validation and prompt-injection defenses that OWASP's security guidance for LLM apps[17] says to build before you expose one publicly. Together they close the demo-to-production gap[18].

> **CORTEXA'S READ**
>
> The demo is the easy part that makes the promise; production is the work that keeps it. Anyone quoting a ship date off a slick demo is selling hype/snake oil.

The "so what" for agencies: when a client greenlights off a demo, reset the expectation kindly — the demo proved it's feasible, not ready. Scope the evals, grounding, guardrails, and review; that's where the timeline and the value actually live.

7. The agency playbook

Put it together and a pattern emerges: almost every AI question a client brings you is really one of six — and each has an honest, scopeable answer. Here's the playbook you can run in a meeting:

- **"Can we train it on our data?"** → a grounding (RAG) project: which documents should it answer from? (Ch. 2)
- **"The AI got it wrong."** → grounding plus a review step — not a better prompt. (Ch. 1)
- **"We need a prompt engineer."** → coaching plus a library of on-brand examples. (Ch. 3)
- **"Can it look at / listen to this?"** → usually yes, the cheap useful kind — scope it by sense and direction. (Ch. 4)
- **"Build us an AI agent."** → a specific, checkable task with a human in the loop. (Ch. 5)
- **"We saw it work — ship it."** → evals, grounding, guardrails, review; feasible isn't ready. (Ch. 6)

The through-line is the one from chapter 1: AI is a confident guesser, so the job is always the same — ground it, check it, and keep a human on the calls that matter. Sell that honest version and you're the partner still standing when the hype cools.

> **CORTEXA'S READ**
>
> Anyone can conjure a demo now. The rare, billable skill is the judgment in this playbook: knowing what's real, what to scope, and what to hand back with your client's name on it. That's the force-multiplier work — and it's exactly what the AI can't do for itself.

Sources

1. Ji et al. — Survey of Hallucination in Natural Language Generation (ACM Computing Surveys)
   https://dl.acm.org/doi/10.1145/3571730
2. Anthropic — Pricing (per-token, input and output)
   https://www.anthropic.com/pricing
3. Cortexa Ground Truth Nº 2 — "AI doesn't lie — it guesses confidently"
   https://www.cortexa-consulting.ai/insights/why-ai-guesses
4. Cortexa Ground Truth Nº 5 — "What a 'token' is, and why it's on your invoice"
   https://www.cortexa-consulting.ai/insights/what-is-a-token
5. Lewis et al. — Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
   https://arxiv.org/abs/2005.11401
6. Cortexa Ground Truth Nº 3 — "RAG is just giving AI an open book"
   https://www.cortexa-consulting.ai/insights/what-is-rag
7. Cortexa Ground Truth Nº 4 — "Fine-tune, retrieve, or just ask better"
   https://www.cortexa-consulting.ai/insights/fine-tuning-vs-rag-vs-prompting
8. Anthropic — Prompt engineering: best practices
   https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/claude-prompting-best-practices
9. Cortexa Ground Truth Nº 6 — "'Prompt engineering' is mostly just clear writing"
   https://www.cortexa-consulting.ai/insights/prompt-engineering-is-clear-writing
10. OpenAI — Hello GPT-4o
    https://openai.com/index/hello-gpt-4o/
11. Cortexa Ground Truth Nº 9 — "What 'multimodal' really means for a campaign"
    https://www.cortexa-consulting.ai/insights/what-multimodal-means
12. Anthropic — Building effective agents
    https://www.anthropic.com/research/building-effective-agents
13. Cortexa Ground Truth Nº 1 — "What 'agentic AI' actually means"
    https://www.cortexa-consulting.ai/insights/what-agentic-ai-actually-means
14. NIST — AI Risk Management Framework
    https://www.nist.gov/itl/ai-risk-management-framework
15. Anthropic — Define success criteria and build evaluations
    https://platform.claude.com/docs/en/docs/test-and-evaluate/develop-tests
16. Cortexa Ground Truth Nº 8 — "Evals, in plain English"
    https://www.cortexa-consulting.ai/insights/evals-in-plain-english
17. OWASP — Top 10 for Large Language Model Applications
    https://genai.owasp.org/llm-top-10/
18. Cortexa Ground Truth Nº 7 — "A great demo isn't a shipped product"
    https://www.cortexa-consulting.ai/insights/demo-to-production-gap