

FIELD GUIDE

Shipping AI, Without the Surprises.

The surprises AI work meets after launch are cheaper to handle at kickoff. Nine explainers, from the labels the law wants to the proof a client asks for.



N° 30

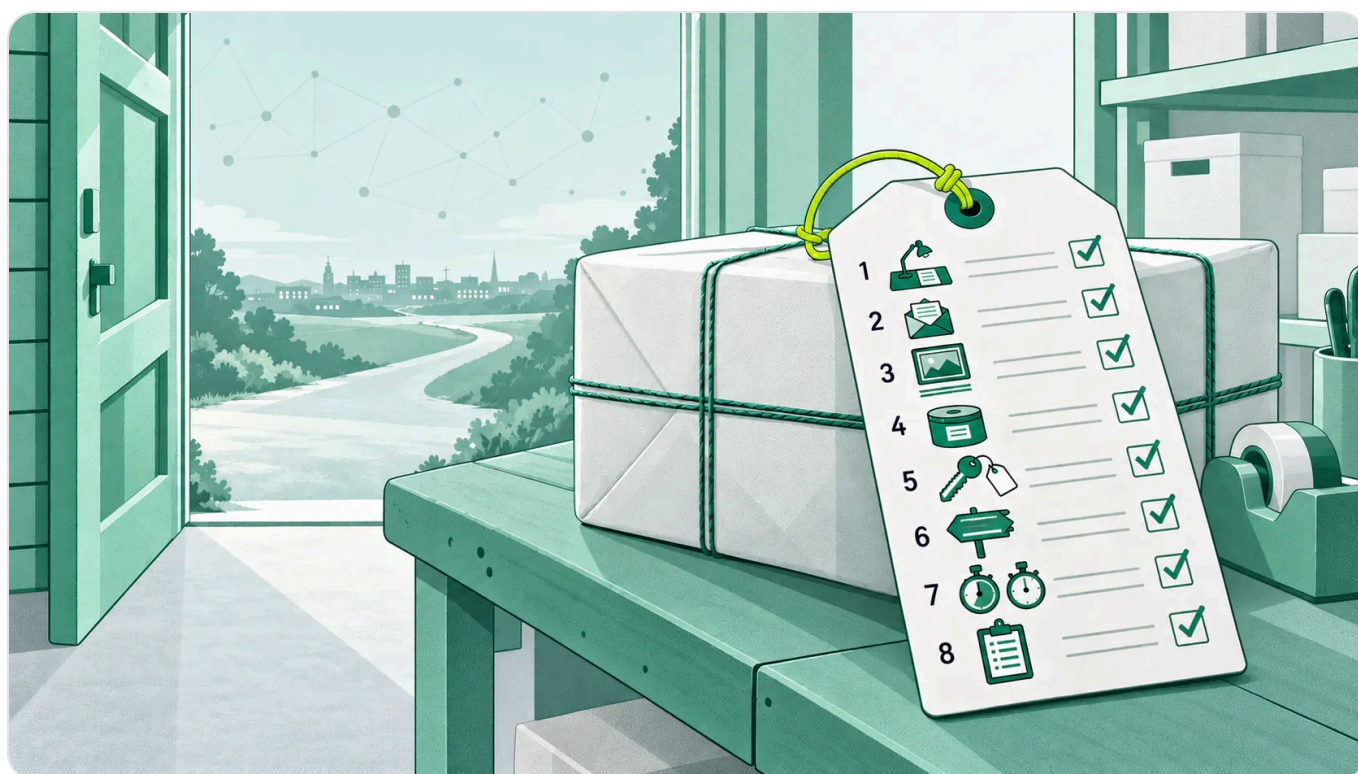
Shipping AI, Without the Surprises.

The Cortexa Field Guide — issues 21 to 29, from launch to proof.

Intelligence, **applied.**

*Intelligence, **applied.*** · October 15, 2026 · 20 min read · www.cortexa-consulting.ai/insights/shipping-ai-without-surprises

The first Field Guide, [AI, Without the Hype¹](#), covered what Artificial Intelligence (AI) is. The second, [Building With AI, Without the Guesswork²](#), covered how to build with it. This third one covers what happens after the work ships, when it meets people who never saw the brief: a regulator, a search page, a shopping assistant, a stranger with bad intentions, and eventually the client's finance team. Nine explainers sit behind these eight chapters. Each opened on a question someone asked out loud, and nearly every one of those questions is cheaper to answer before launch than after. So read this as the list to run through at kickoff. Each chapter names the surprise, shows what the evidence supports, and points back to the explainer that has the detail.



The eight chapters as one shipping tag. Each row gets its tick before the work leaves the bench.

1. What goes in

Two choices shape how an AI feature behaves long before anyone outside sees it: which model runs it, and what that model gets to read. Both reward the same instinct. Start small. The first is [size⁴](#), and inside one model family the price gap is already five to one: Anthropic lists its smallest Claude model at [one dollar per million input tokens](#) and its largest at [five¹²](#). The second is [context⁷](#), meaning the instructions, documents, tools and history the model sees when it answers. Anthropic treats that as [a finite resource with diminishing marginal returns¹³](#). Pour in everything and the answer can get worse.



A cleared desk helps the model find the page that matters.

- Small models have closed a surprising amount of ground. The smallest model to clear 60% on a broad knowledge test went from 540 billion parameters in 2022 to 3.8 billion in 2024, a 142-fold reduction¹⁴.
- Trying the cheap model first is a proven design. Stanford's FrugalGPT sent each query to cheaper models before escalating, and matched the best single model with up to a 98% cost reduction¹⁵.
- Where a fact sits changes whether it gets used. Models handle information at the start and end of a long input better than in the middle, even for explicitly long-context models¹⁶. A key fact buried halfway down can be walked straight past.

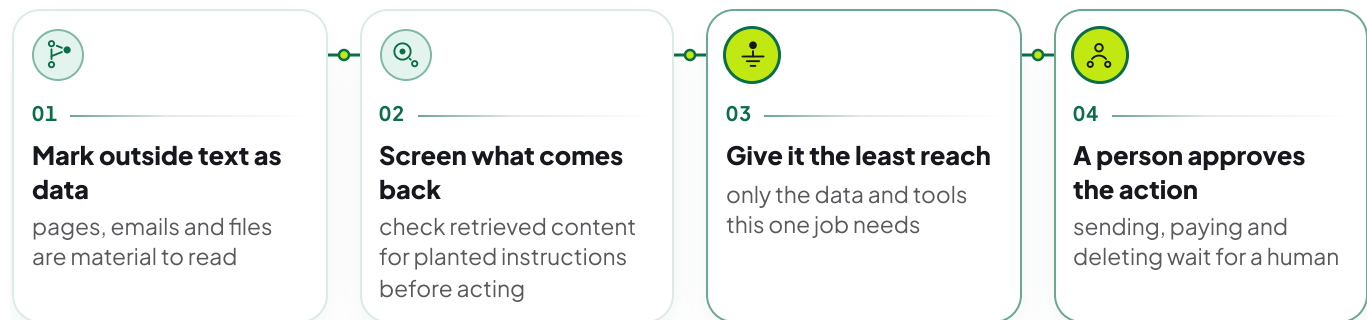
CORTEXA'S READ

When a feature misfires, the reflex is to buy horsepower. Look at what the current model was handed first: the stale document, the ten tools where two would do, the one fact buried in the middle. Fix the feeding, measure again on the client's real examples, and only then decide whether the bigger model earns its price.

2. Who gives the orders

An AI feature that reads outside content can be steered by it. Prompt injection¹¹ is the name for instructions planted in that content, and Anthropic describes them as adversarial instructions hidden within the content that AI models process¹⁷. The Open Worldwide Application Security Project (OWASP) puts prompt injection¹⁸ first on its list of risks for apps built on a Large Language Model (LLM). The reason it works ties back to chapter one. Your instructions, a retrieved web page and a customer's email all arrive in the same stream of text, so a sentence that sounds like an order can be taken as one, whoever wrote it.

DESIGNING FOR THE INJECTION THAT GETS THROUGH



🔄 Plan for one layer failing. The next one is there for exactly that.

- Nobody claims a cure. OpenAI writes that prompt injection is unlikely to ever be fully “solved”¹⁹, and OWASP says no reliable prevention mechanism exists today¹⁸.
- Defenses cut the odds. None closes them. In Anthropic's own browser-agent testing, an attacker still got through about 1% of the time against its latest model and safeguards, a result the company says still represents meaningful risk¹⁷.
- The worst case depends on reach. Simon Willison's lethal trifecta²⁰ is private data, untrusted content and a way to send things out. Take one of the three away and that particular path to stolen data is gone.

CORTEXA'S READ

Assume some planted instruction will get through, and design so the worst it can produce is an odd draft a person catches. Decide what the feature is allowed to touch while it is still a line in the proposal. At that stage the decision is free.

3. Saying it was AI

The first surprise from outside is legal. In 2026 three governments wrote AI disclosure into law, and the [labeling explainer](#)³ maps all three. The European Union (EU) AI Act's transparency rules apply from [August 2, 2026](#)²¹: providers must mark generated content so it is [detectable as artificially generated](#)²², and deployers must [clearly label](#)²² deepfakes. California's AI Transparency Act reaches generative systems with [over 1,000,000 monthly](#)²³ users. New York now requires an advertiser who knows an ad contains a [synthetic performer](#)²⁴ to [conspicuously disclose](#)²⁴ it. None of this replaces counsel on a live campaign.



Credit the machine the way you already credit a photographer: one small line under the picture.

- Most agencies use these tools; few build them. California's duties to label output and offer a free detector fall on the provider of a system that size, which is usually the tool vendor.
- The fines are modest. New York's runs \$1,000 for a first violation and \$5,000 for each after, per its [synthetic performer law](#)²⁴. The bigger bill is re-cutting an ad after it has already run.
- Key words are still loose. New York leaves "conspicuous" undefined, and the EU's "clearly" will be settled by [Commission guidelines](#)²¹ and early practice.

CORTEXA'S READ

Treat the label as a credit line. Ask two things when the brief lands: which parts will a machine make, and which markets will see them? Those answers pick the rule while changing course is still cheap.

4. Proving where it came from

A label tells the audience AI was involved. A [content credential](#)⁵ is the record underneath it: signed metadata in the file saying who made it, when, with what tool, and whether AI generated or altered it. The open standard is the Coalition for Content Provenance and Authenticity (C2PA). The big generators already write provenance by default. OpenAI [attaches C2PA Content Credentials to the images its models generate](#)²⁵, and Google embeds a [SynthID watermark](#)²⁶ in the images, audio and video its models make.



What is in the file never changes. The record of where it came from is the part that goes missing.

- A signature makes tampering visible. Change the file and the signed record stops matching, so a checker can flag it.
- Metadata falls off in transit. A screenshot or a re-encoding upload can strip the credential while the picture looks identical. Watermarks written into the pixels are the paired method,

and SynthID is built to survive cropping, filters, and compression²⁶.

- The National Institute of Standards and Technology (NIST) treats provenance, watermarking and detection as complementary approaches to content transparency²⁷. Each fails in a different place, which is the argument for using more than one.

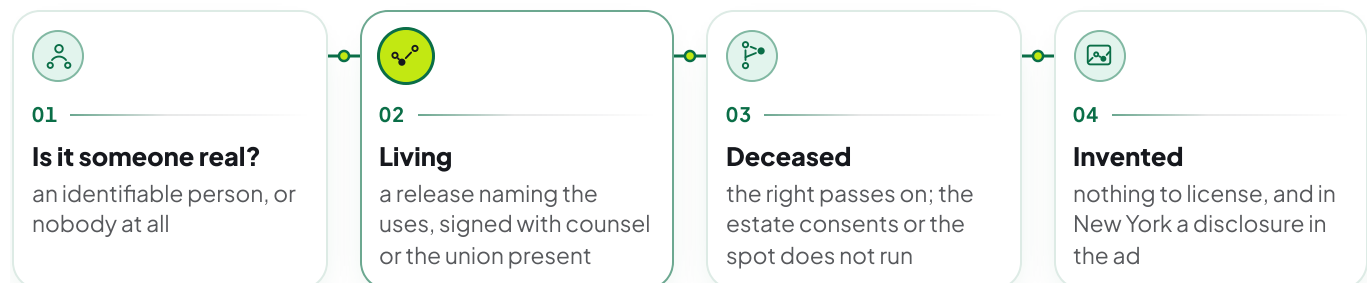
CORTEXA'S READ

Use credentials to prove things about your own work: this came from us, on this date, with AI used here and not there. A missing credential proves nothing about anyone else's file, because it may have been stripped or never written. Pitch it to a client as a receipt. Keep your signed originals on file, since that copy survives when the social upload does not.

5. The face and the voice

Cloning a face or a voice now takes minutes of footage. Whether you may run the result in an ad is a question the software never asks, and California and New York now answer it in statute, as the likeness explainer⁹ sets out. Both states went after the contract. In California a digital-replica clause is unenforceable when it does not include a reasonably specific description of the intended uses²⁸ and the performer had neither a lawyer nor a union agreement covering replicas. New York voids the same kind of clause²⁹ on the same conditions. What counts as a replica is whether an audience would take it for the person: California asks if it is readily identifiable²⁸, and New York whether a layperson would not be able to readily distinguish²⁹ it from the real one.

CLEARING AN AI FACE OR VOICE



🔄 *The copy is easy in every case. The signature is what makes it usable.*

- Death does not free the likeness. An unauthorized replica of a deceased personality costs the greater of ten thousand dollars (\$10,000) or the actual damages suffered³⁰ in

California, and the greater of two thousand dollars or the compensatory damages³¹, plus profits, in New York.

- An invented spokesperson has nobody to license, and it brings a disclosure duty. New York's advertiser must conspicuously disclose²⁴ the synthetic performer in the ad itself.
- This is state law. A national campaign reaches California and New York by default, so the strictest rule your media plan touches is the one to work to.

CORTEXA'S READ

Put likeness on the same clearance sheet as music and stock footage, with named uses, a term, a territory and a signature. If the honest answer to “whose face is this?” is “we generated it,” the disclosure question still needs an answer before the cut is locked.

6. When machines read you first

More and more, the first audience for a client's work is software. On Google, an AI Overview now answers many questions above the links, written from the same pages that rank⁶. Google says its AI features rely on our core Search ranking systems to retrieve relevant, up-to-date web pages³². A Pew Research Center study found that a search showing a summary led to a click on a traditional result in 8% of visits, against 15% when no summary appeared³³. A page can hold first place and still lose the visit.

Shopping is heading the same way. When a shopper asks an assistant what to buy, what gets picked⁸ depends on what the model can read. OpenAI says ChatGPT's product results are organic and unsponsored³⁵, surfaced when the model perceives it to be relevant to your intent³⁶ and assembled from structured metadata³⁶: the feeds, catalogs and pages it can parse. A product whose data the model cannot read has little chance of making that list, however loud its advertising.



The sign answers the question on the road. The visitors who still walk in are the ones who came to buy.

- Google disputes the traffic story. Its Search lead says total organic clicks are relatively stable year-over-year³⁴, without publishing figures. Weigh that against the independent counts before a forecast rests on it.
- Clean product data is the price of entry. OpenAI asks merchants to share your product feed³⁷, and Google added dozens of new data attributes³⁸ to Merchant Center for discovery.
- Paid placement is coming in its own lane. Google is continuing to test ads in AI Mode³⁸, kept apart from the organic results, so plan for a paid track beside the earned one.

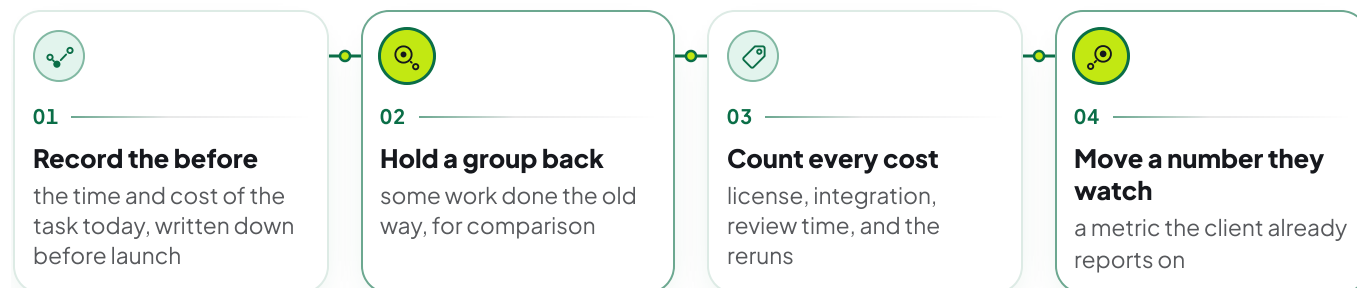
CORTEXA'S READ

Search and shopping have converged on one job: be the clearest thing a machine can read. Put the answer where a summary can lift it, keep the catalog complete and current, and change the scoreboard before the client sees the graph. Informational sessions will slide even when the work is sound. Citations in AI answers, branded demand and sales are what still tell the truth.

7. Proving it paid off

Sooner or later the client asks whether it was worth it. What holds up is a number measured against what came before¹⁰, after the running cost comes out, and it is usually smaller than the vendor's slide. When Stanford's AI Index looked at what companies report, firms that saw savings mostly reported cost savings of less than 10%³⁹. Where revenue rose, the most common level of revenue increases is less than 5%³⁹.

BEFORE YOU CLAIM THE WIN



If you can't trace the gain to the tool after its cost, don't promise it.

- On well-suited work the lift is real. In a Boston Consulting Group (BCG) field experiment, consultants using AI completed tasks 25.1% more quickly⁴⁰ and produced 40% higher quality results⁴⁰ than a group working without it.
- On the wrong work it backfires. Given a problem built to sit outside the model's competence, the consultants who leaned on AI did worse⁴⁰ than those who did not.
- Measurement is a schedule. NIST's risk framework says AI systems should be tested before their deployment and regularly while in operation⁴¹.

CORTEXA'S READ

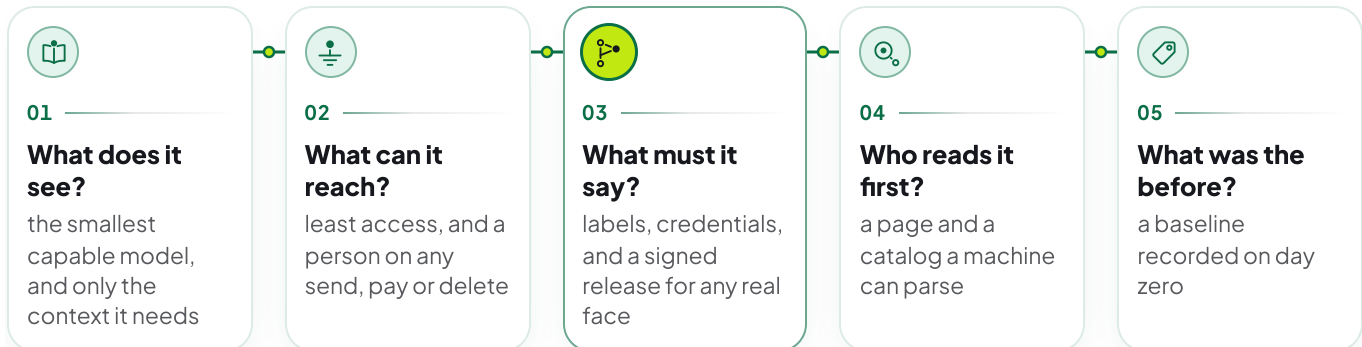
Write the baseline down on day zero. It costs nothing, and without it you can only claim the lift. Then offer the client a metric they already watch and a date to read it. A promise with a date on it is one you can be held to, which is why it persuades.

8. The agency playbook

Seven chapters share one pattern: every surprise in them is cheaper to handle before launch than after. So the playbook moves the questions to the front, in the order the work meets the world. Decide what the model sees and how big it needs to be. Limit what a fooled feature can

reach. Settle labels, credentials and consent while the edit is still open. Make the work legible to the machines that read it first. And write down the before, so the result can be shown.

THE QUESTIONS TO ASK BEFORE IT SHIPS



🔄 *Change the feature and every question after it opens again.*

- Write disclosure and likeness into the statement of work (SOW), with a named approver for each. Flagging the exposure is your job. Deciding is the client's, with their counsel.
- Ask every vendor the same short list: which model runs where, what outside content the feature reads, what it can do if that content fools it, and which steps a person approves.
- Report what still carries intent. Informational sessions will fall. Citations, branded demand and conversions tell the client how the work is doing.
- Book the measurement as part of the job, with a baseline before build and a reading date after launch. It turns “we think it helped” into a number the client can take to their own boss.

CORTEXA'S READ

None of this slows the work much. It moves a handful of questions from the post-mortem to the kickoff, where the answers are cheap. That shift does more than any tool choice to keep a client's name and your reputation out of the surprises.

Sources

1. Cortexa Ground Truth N° 10 — “AI, Without the Hype — The Cortexa Field Guide”
<https://www.cortexa-consulting.ai/insights/whats-real-in-ai>
2. Cortexa Ground Truth N° 20 — “Building With AI, Without the Guesswork.”
<https://www.cortexa-consulting.ai/insights/building-with-ai-without-guesswork>
3. Cortexa Ground Truth N° 21 — “Labeling AI content just became law.”
<https://www.cortexa-consulting.ai/insights/labeling-ai-generated-content>
4. Cortexa Ground Truth N° 22 — “You rarely need the biggest model.”
<https://www.cortexa-consulting.ai/insights/right-sizing-ai-models>
5. Cortexa Ground Truth N° 23 — “A content credential is a nutrition label for your media.”
<https://www.cortexa-consulting.ai/insights/content-credentials-c2pa>
6. Cortexa Ground Truth N° 24 — “You can rank first and still lose the click.”
<https://www.cortexa-consulting.ai/insights/ai-overviews-and-organic-traffic>
7. Cortexa Ground Truth N° 25 — “Prompt engineering was step one. Now engineer the context.”
<https://www.cortexa-consulting.ai/insights/context-engineering>
8. Cortexa Ground Truth N° 26 — “When an AI shops, what gets picked?”
<https://www.cortexa-consulting.ai/insights/getting-picked-by-ai-shopping-agents>
9. Cortexa Ground Truth N° 27 — “An AI version of a real person still needs consent.”
<https://www.cortexa-consulting.ai/insights/ai-likeness-and-voice-rights>
10. Cortexa Ground Truth N° 28 — “To prove the AI paid off, measure the baseline.”
<https://www.cortexa-consulting.ai/insights/measuring-ai-roi>
11. Cortexa Ground Truth N° 29 — “Prompt injection: who’s giving your AI its orders?”
<https://www.cortexa-consulting.ai/insights/prompt-injection>
12. Anthropic — Claude model pricing
<https://claude.com/pricing>
13. Anthropic — “Effective context engineering for AI agents”
<https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents>
14. Stanford HAI — 2025 AI Index Report, Technical Performance
<https://hai.stanford.edu/ai-index/2025-ai-index-report/technical-performance>
15. Chen, Zaharia & Zou — “FrugalGPT” (preprint)
<https://arxiv.org/abs/2305.05176>
16. Liu et al. — “Lost in the Middle: How Language Models Use Long Contexts” (preprint)
<https://arxiv.org/abs/2307.03172>
17. Anthropic — “Mitigating the risk of prompt injections in browser use” (November 24, 2025)
<https://www.anthropic.com/research/prompt-injection-defenses> — The 1% attack-success figure is vendor-stated, measured by Anthropic against its own internal adaptive attacker.

18. OWASP — LLM01:2026 Prompt Injection (Top 10 for LLM Applications 2026)
https://github.com/GenAI-Security-Project/GenAI-LLM-Top10/blob/main/2026/final/LLM01_PromptInjection.md
19. OpenAI — “Continuously hardening ChatGPT Atlas against prompt injection attacks”
<https://openai.com/index/hardening-atlas-against-prompt-injection/>
20. Simon Willison — “The lethal trifecta for AI agents: private data, untrusted content, and external communication”
<https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/>
21. European Commission — Quick Facts: Transparency rules for AI systems
<https://digital-strategy.ec.europa.eu/en/factpages/quick-facts-transparency-rules-ai-systems>
22. European Commission — Transparency obligations under Article 50 of the AI Act (FAQ)
<https://digital-strategy.ec.europa.eu/en/faqs/transparency-obligations-under-article-50-ai-act>
23. California Legislature — SB 942, California AI Transparency Act (bill text)
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB942
24. New York State Senate — S8420A, synthetic performer disclosure (bill text)
<https://www.nysenate.gov/legislation/bills/2025/S8420/amendment/A>
25. OpenAI — Advancing content provenance (C2PA Content Credentials)
<https://openai.com/index/advancing-content-provenance/>
26. Google DeepMind — SynthID
<https://deepmind.google/models/synthid/>
27. NIST — Reducing Risks Posed by Synthetic Content (AI 100–4)
<https://www.nist.gov/publications/reducing-risks-posed-synthetic-content-overview-technical-approaches-digital-content>
28. California Legislature — AB 2602 (2024), digital replicas in performer contracts (Labor Code § 927)
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2602
29. New York State Senate — General Obligations Law § 5–302, digital replicas in contracts
<https://www.nysenate.gov/legislation/laws/GOB/5-302>
30. California Legislature — AB 1836 (2024), digital replicas of deceased personalities (Civil Code § 3344.1)
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB1836
31. New York State Senate — Civil Rights Law § 50–f, deceased performers' digital replicas
<https://www.nysenate.gov/legislation/laws/CVR/50-F>
32. Google Search Central — Optimizing for generative AI features on Google Search
<https://developers.google.com/search/docs/fundamentals/ai-optimization-guide>
33. eMarketer — Google’s AI summaries are ending search sessions, reducing click-through behavior (reporting Pew Research Center)
<https://www.emarketer.com/content/google-ai-summaries-ending-search-sessions-reducing-click-through-behavior>
34. Google — Liz Reid, “AI in Search is driving more queries and higher quality clicks”
<https://blog.google/products/search/ai-search-driving-more-queries-higher-quality-clicks/>
35. OpenAI — “Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol”
<https://openai.com/index/buy-it-in-chatgpt/>

36. OpenAI Help Center — “Shopping with ChatGPT search”
<https://help.openai.com/en/articles/11128490-improved-shopping-results-from-chatgpt-search>
37. OpenAI — “Power product discovery in ChatGPT”
<https://openai.com/chatgpt/search-product-discovery/>
38. Google — “New tech and tools for retailers to succeed in an agentic shopping era”
<https://blog.google/products/ads-commerce/agentic-commerce-ai-tools-protocol-retailers-platforms/>
39. Stanford HAI — 2025 AI Index Report, Economy
<https://hai.stanford.edu/ai-index/2025-ai-index-report/economy>
40. Ethan Mollick, One Useful Thing — “Centaur and Cyborgs on the Jagged Frontier” (reporting the Dell’Acqua et al. BCG field experiment)
<https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged>
41. NIST — AI Risk Management Framework (AI RMF 1.0), Core: the MEASURE function
<https://airc.nist.gov/airmf-resources/airmf/5-sec-core/>