
They Dropped the Exit Fee

Continual learning was supposed to give AI labs their moat. The contracts already say otherwise, and the last three platform transitions already told us where the switching cost lives.

Gerald C. Kane, Ph.D.

C. Herman and Mary Virginia Terry Chair in Business Administration
Department Head, Management Information Systems
Terry College of Business, University of Georgia

© 2026 Gerald C. Kane. Licensed under CC BY-NC 4.0: share and adapt for noncommercial use with credit to the author; do not present as your own. Full terms on the last page.

Executive Summary

OpenAI's developer documentation carries a maintenance notice that almost no enterprise buyer will ever read. On May 7, 2026, the company notified developers of an update to availability: "Creating fine-tuning jobs or training is not available to organizations that have not previously run fine-tuning." Two more dates follow on the same page. On July 2, 2026, the door closed on organizations that had not run inference on a fine-tuned model in the previous sixty days. On January 6, 2027, it closes on everyone: "Active existing customers will no longer be able to create new fine-tuning jobs on this date" (OpenAI, 2026a). Three months after that first notice, a widely read essay argued that the next phase of the industry belongs to models that update their own weights from deployment traffic, and that this will finally give the labs the moat they lack, because a model that has learned your organization becomes too expensive to abandon (Patel, 2026). OpenAI is withdrawing the mechanism that prediction requires from the market, on a published schedule, in writing.

The temptation is to treat continual learning as the development that changes the terms of the enterprise software relationship. It does not change them. The lock-in question is old. We have run it through enterprise resource planning, through electronic health records, and through public cloud, and the answer has held each time: the switching cost that appears in the contract is the small half of the problem, and the switching cost that appears in the organization is the large half. Nothing in the current technical record suggests this round will be different, and quite a lot in the current commercial record suggests it will be the same.

Begin with the market evidence. No provider currently sells continual updating of a customer-specific base model from production traffic, and the model layer moves freely without it: Menlo Ventures' enterprise surveys record OpenAI's share of enterprise large-language-model spend falling from 50 percent in 2023 to 27 percent in 2025 while Anthropic's rose from 12 to 40, with organizations typically routing across three or more models (Menlo Ventures, 2024, 2025). A moat that does not exist cannot be what holds customers in place. What binds is the workload: evaluation suites written against one model's behavior, prompt and tool scaffolding tuned to one provider's quirks, security review already cleared, data-processing agreements already signed, and staff who have learned one system's failure modes. Those costs live at the level of a validated production deployment, and none of them appears in the contract.

The bargain the prediction anticipates already exists in operative contract text, and it runs the other direction. Google's Gemini API Additional Terms of Service, effective March 23, 2026, split training rights by payment status inside the agreement itself: content submitted to unpaid tiers is used to develop Google products and machine learning technologies, with human reviewers who "may read, annotate, and process your API input and output," while paid tiers are exempt (Google, 2026). OpenAI runs a live programme granting complimentary daily tokens in exchange for shared traffic, and states that "this setting is not available to certain organizations, including Enterprise and customers with Zero Data Retention enabled" (OpenAI, 2026c). Enterprises are not being paid to let a model learn them. They are paying for exemption.

The historical case is cleaner than any forecast. The most-cited technical lock-in of the cloud era was the egress fee. Article 29 of the EU Data Act caps switching charges at directly incurred cost

from January 11, 2024 and prohibits them outright from January 12, 2027 (European Union, 2023). Google Cloud removed exit egress charges in January 2024 and AWS followed in March 2024 (Google Cloud, 2024; Amazon Web Services, 2024). The contractual half of cloud lock-in was dismantled in roughly twenty-four months, and enterprises still largely stayed where they were. When you remove the contractual cost and the customers do not move, the cost was never contractual.

This paper argues that the enterprise decision implied by the continual-learning thesis, signing away training rights now to secure a future capability advantage, is not supported by anything currently purchasable. It sets out what is actually shipping, what the four major providers' contracts actually say and how much they differ, what the egress and ERP record establishes about where switching costs live, and what a board should do when a lab offers to learn its organization.

Why This Matters Now: Four Curves, and the One That Moved First

The Technology Fallacy (Kane et al., 2019) put forward an image we have used more than any other in the years since: technology, individuals, organizations, and public policy all change at different rates. Technology moves fastest. Individuals adapt more slowly, organizations more slowly still, and public policy always lags. The organization is the one that has to absorb the mismatch, which is why organizational problems are what technology transitions actually produce.

The image is doing something unusual in this case, and the unusual part is what makes the argument. Public policy moved first.

The EU Data Act's switching-charge provisions were adopted in 2023 and began binding in January 2024, before the AI procurement question was fully formed and years before continual learning became a commercial talking point. Two hyperscalers dropped exit egress charges ahead of the deadline. The contractual constraint on leaving was removed by regulation and competitive response acting together, faster than anyone predicted, and the organizational constraint on leaving did not move at all. That sequence is the cleanest natural experiment the platform era has given us on where lock-in lives, and it produced an unambiguous answer.

The second construct this paper puts under load is one from our prior work in this series, which argued that ownership of an AI capability is a portability discipline rather than a possession: you own the loop, not the model, and you keep ownership by keeping the ability to move. That framing survives the evidence here, but only barely, and it needs a correction. Portability discipline lives mostly outside the contract. Contracts govern the small half of the problem. The evaluation suite you can run against a new provider on Tuesday morning, the retrieval index that lives in your own storage rather than in a provider-specific store, the record of which prompts broke and why, the fact that three people on your team have actually shipped against two providers rather than one: those are the portability discipline, and none of them appear in a master services agreement.

Everything that follows is an attempt to say precisely which half of the problem the continual-learning debate is arguing about, and to point the reader at the other half.

I. Three Different Things Are Called "The Model Learns Your Business"

The category confusion in this discussion is worth resolving before anything else, because all three tiers are marketed with the same sentence, only one of them touches weights at all, and none touches base weights.

The first tier is retrieval and memory over a frozen model, which is what nearly everyone actually means. Every major provider ships it. Microsoft's Foundry documentation is unusually direct about the architecture: "The models are stateless: no prompts or completions are stored in the model. Additionally, prompts and completions are not used to train, retrain, or improve the base models" (Microsoft, 2026). Stateful behavior comes from separate data stores that persist message history alongside the model. Grounding features work the same way, retrieving from a connected source and augmenting the prompt. Nothing in this tier creates lock-in at the level of weights. It creates lock-in at the level of a data store, which is a materially different and far more portable thing, and which most enterprises could unwind in a quarter if they had ever written down what was in it.

The second tier is per-customer adapter tuning, and it is real, shipping, and receding at the largest provider. Google's Vertex supervised tuning for Gemini exposes an `adapter_size` parameter controlling low-rank adapter rank, which confirms the mechanism is a small adapter over a shared frozen base rather than a customer-specific model (Google Cloud, n.d.-b). OpenAI states that fine-tuned models are "for your use alone and never served to or shared with other customers or used to train other models" (OpenAI, 2026b). Microsoft says the same: "Your fine-tuned Models sold by Azure are available exclusively for your use" (Microsoft, 2026). This tier does create a genuine, if modest, technical switching cost. It is also the tier OpenAI is winding down, which is the subject of the next section.

The third tier, continual updating of base weights from deployment traffic, is not commercially available from any major provider. What exists instead is a research literature, and that literature is candid about why. The clearest demonstration is SEAL, published at NeurIPS 2025, in which a model generates its own "self-edits" that are applied as gradient updates through an inner loop while an outer reinforcement-learning loop optimizes the edit generation. The reported results are real: question-answering accuracy improving from 32.7 percent with no adaptation to 47.0 percent after two rounds in the single-passage setting, and 72.5 percent on a simplified ARC subset against zero percent for in-context learning (Zweiger et al., 2025). The authors also report the limitation that blocks deployment. Repeated self-edits degrade performance on earlier tasks. Google Research frames continual learning as an open architectural problem requiring a new multi-frequency memory design, which is a statement that scaling the current stack does not solve it (Google Research, 2025). The empirical baseline is broader still: catastrophic forgetting is observed generally in models from one billion to seven billion parameters during continual fine-tuning (Luo et al., 2025).

There is also a hard economic reason the third tier collapses into the second. Serving many customer-specific adapters over one shared base is a solved engineering problem. Serving many customer-specific *base models* is not, because it destroys the batching economics that sparse models depend on. DeepSeek's operator disclosure remains the only public accounting of frontier-scale serving economics, and it states the mechanism plainly: with only 8 of 256 experts active per

layer, "the model's high sparsity necessitates an extremely large overall batch size." Its disclosed decode configuration spanned eighteen nodes, 144 H800 GPUs, at an average of 226.75 nodes and \$87,072 per day across the measured window (DeepSeek, 2025). A specific concurrency threshold has circulated in commentary as the figure at which per-customer weight forks become affordable. That number is not publicly established, and it does not appear in the DeepSeek disclosure or in any other primary operator source we located. The defensible version of the claim is stronger anyway: efficient sparse-model serving is organized around deployment units of scores to hundreds of GPUs, those units only pay for themselves when kept full, and a per-customer base-model fork is a commitment to keep a unit full for one customer.

The implication for leaders. When a vendor says the model will learn your business, ask which of the three tiers is being described, and ask it in writing. If the answer is memory and retrieval, the asset is your data store and you should know where it physically lives. If the answer is adapter tuning, ask what happens to the adapter when the base model is deprecated. If the answer is the third tier, ask for the contract terms, because as of this writing there are none to show you.

II. OpenAI Is Retiring the Mechanism the Thesis Requires

The single most useful fact in this entire debate is a published schedule, and it points the opposite way from the forecast.

OpenAI's deprecations page lays out the self-serve fine-tuning wind-down in three dated steps. From May 7, 2026, fine-tuning is closed to organizations that have never used it. From July 2, 2026, it is closed to organizations that have not run inference on a fine-tuned model in the past sixty days. From January 6, 2027, no customer can create new fine-tuning jobs, with inference on existing fine-tuned models continuing only until the underlying base model is itself deprecated (OpenAI, 2026a). The stated rationale, that current base models follow instructions well enough that prompting plus retrieval is cheaper and sufficient, has been widely reported but we did not locate it as a verbatim company statement, so it should be treated as reported rather than quoted. The rationale matters less than the behavior. In the same season an essay predicted per-customer weights would become the industry's moat, OpenAI made per-customer weight adaptation harder to buy.

The same page contains a second fact that reframes the whole switching-cost conversation. On October 23, 2026, twelve legacy models and snapshots shut down at once, including gpt-3.5-turbo, gpt-4-0613, gpt-4-turbo, gpt-4o-2024-05-13, o1, o1-pro, o3-mini, and o4-mini, nearly all with a designated substitute in the gpt-5.6 family. The Evals API and dashboard are scheduled to shut down on November 30, 2026, going read-only a month earlier (OpenAI, 2026a).

Put those two facts together. The harder problem for an enterprise in 2026 is staying on the model it validated. The compliance artifact that regression testing, model documentation, and audit all depend on is a frozen checkpoint, and providers are retiring frozen checkpoints on twelve-month cycles while the organizations that certified against them are still working through their second annual review.

This is where the continual-learning thesis has a real problem that its critics have mostly missed. A model that updates continuously does not merely add a switching cost. It removes the fixed artifact against which anything can be certified at all. The AI governance apparatus that regulators,

insurers, and internal audit have spent three years building assumes a model that stops changing long enough to be examined. A continually adapting model breaks that assumption, and it breaks it for the enterprise before it breaks it for the lab.

The implication for leaders. Put your provider's deprecation calendar into the same governance cycle as your evaluation results. If a model your organization certified is scheduled to shut down inside your certification window, that is a finding for the audit committee, and it is a more immediate finding than any lock-in scenario in this paper.

III. Where the Enterprise Record Says the Switching Cost Sits

Menlo Ventures' enterprise surveys are the most-cited spend measurements in this market, and the pattern they document is movement. Stating that record carefully matters, because it decides where the switching cost can and cannot sit.

The headline numbers first, since they establish the sample is worth taking seriously. Menlo reports enterprise generative AI spend of \$37 billion in 2025, up from \$11.5 billion the year before, on a survey of 495 U.S. enterprise decision-makers. It estimates Anthropic at 40 percent of enterprise large-language-model spend, up from 24 percent the prior year and 12 percent in 2023, with OpenAI at 27 percent, down from 50 percent in 2023. It reports enterprise open-source model share falling from 19 percent to 11 percent year over year (Menlo Ventures, 2025).

The prior year's survey states the mechanism behind those share shifts plainly. Organizations typically deploy three or more foundation models and route among them by use case. OpenAI's erosion happened because, in Menlo's words, "some enterprises switched from GPT-4 to Claude 3.5 Sonnet when the new model became state-of-the-art," and enterprises moving to a new model most commonly cited security and safety, price, performance, and expanded capabilities as their reasons. The same survey found only 9 percent of production models were fine-tuned at all (Menlo Ventures, 2024). This is a market in which share reallocates when capability moves, on evaluation criteria enterprises can name.

Hold that record next to the technical one from Section I, and the distinction that matters comes into focus. There is no weight-level adaptation to any enterprise's data available for purchase today, and reallocating new spend to a better model is evidently cheap, which is why the share numbers can move by double digits in two years. Moving a validated production workload is what is expensive, and the expense is not hard to locate. It sits in the evaluation suite the team wrote against one provider's behavior and has never rewritten. In the prompt library, the tool definitions, the retry logic, and the retrieval plumbing tuned to one provider's quirks. In the security review that took eleven weeks and cleared exactly one vendor. In the data-processing agreement, negotiated by a legal team with no appetite for doing it twice. In the four engineers who know which prompts break this model and have no idea which prompts break the other one. Every item on that list is an organizational asset, and every item on it would still exist if every model on the market were frozen forever.

There is a structural fact that cuts even harder. The dominant enterprise consumption channel for frontier models is a cloud intermediary, not the lab. AWS describes a deployment architecture in which the model provider has no access to Bedrock logs or to customer prompts and completions

(Amazon Web Services, n.d.). Microsoft states that customer prompts, completions, embeddings, and training data are not available to OpenAI or to other providers of models sold by Azure (Microsoft, 2026). Anthropic's own consumer-terms notice confirms the split from the other side, listing its API, Amazon Bedrock, and Google Cloud's Vertex API among the services to which its consumer training change does not apply (Anthropic, 2025b). If the lab cannot observe the traffic, the lab cannot learn from it, and the moat cannot form through that channel regardless of what the architecture becomes capable of.

The implication for leaders. Run the exercise rather than debating it. Ask your team how long it would take to run your top twenty production prompts against a second provider and score the results against the same rubric. If the honest answer is more than two weeks, you already have the lock-in problem, and continual learning has nothing to do with it.

IV. The Bargain Already Exists in Writing, and Enterprises Are on the Paying Side

This is the finding that reframes the debate, and it comes entirely from operative contract text rather than from anyone's forecast.

Google's Gemini API Additional Terms of Service, effective March 23, 2026, contain the clearest price-conditioned training clause we located anywhere in the market. For unpaid tiers, including Google AI Studio and the unpaid Gemini API quota, the terms state that "Google uses the content you submit to the Services and any generated responses to provide, improve, and develop Google products and services and machine learning technologies, including Google's enterprise features, products, and services, consistent with our Privacy Policy." The same section states that "human reviewers may read, annotate, and process your API input and output," and instructs users in the document's own emphasis: "Do not submit sensitive, confidential, or personal information to the Unpaid Services." For paid tiers, the same document states that "Google doesn't use your prompts (including associated system instructions, cached content, and files such as images, videos, or documents) or responses to improve our products" (Google, 2026).

That is precisely the structure the continual-learning thesis predicts. Free access in exchange for training rights, the search business model applied to inference. It already exists, it is written into a terms-of-service document rather than promised in a keynote, and it is aimed at individual developers and free-tier users. The enterprise side of the same agreement is the opt-out, and the opt-out is what money buys.

OpenAI runs the same coupling in a different shape and states the enterprise exclusion outright. Its data-sharing programme offers that "some organizations may qualify for daily complimentary tokens on traffic shared with OpenAI," with eligible Tier 3 through 5 organizations receiving up to one million tokens per day across the flagship model group and up to ten million per day across the small-model group, and Tier 1 and 2 organizations receiving 250,000 and 2.5 million respectively. A second price coupling sits alongside it: "when sharing data for reinforcement fine-tuning jobs, the resulting models will be billed at discounted inference rates," and shared evaluations run free for up to seven runs per week. Then the exclusion, stated twice. "This setting is not available to certain organizations, including Enterprise and customers with Zero Data Retention enabled." And

separately: "Accounts with Zero Data Retention enabled cannot opt in to our data sharing mechanisms" (OpenAI, 2026c).

The one provider running a live data-for-compute subsidy has built a wall around exactly the population the thesis says will be courted with it.

Anthropic supplies the third data point, and it is the cleanest proof that price and training rights are already coupled. Its August 2025 consumer-terms notice moved Free, Pro, and Max plans to opt-in training, and stated that "we are also extending data retention to five years, if you allow us to use your data for model training," with customers who decline continuing on the prior thirty-day retention period. The notice states explicitly that the change does not apply to services governed by the Commercial Terms, naming Claude for Work, Claude Gov, Claude for Education, and API access including through Bedrock and Vertex (Anthropic, 2025b). One company, two customer populations, and a training-rights term that varies by which one you are in.

We found no instance of the enterprise-facing version of this bargain. No provider located in this research offered an enterprise a discount specifically in exchange for training rights on its work sessions. That absence is the finding, and the plausible reason is not technical. Enterprise contracts, data-processing agreements, business-associate agreements, and zero-data-retention commitments are what make enterprise revenue sellable in regulated industries. A lab that begins training on enterprise work sessions puts its own indemnity and compliance position at risk. That is a contractual constraint on the moat rather than a capability constraint, and contractual constraints do not dissolve because an architecture changes.

The implication for leaders. Your organization is already paying a price for training exemption. It is embedded in your rate card and you almost certainly cannot see it as a line item. The board-level question therefore arrives well before any future data-for-capability offer does. Ask what your organization pays today not to be learned from, whether that is the right trade for each workload, and whether the same terms should govern your customer-service transcripts and your merger analysis. For a meaningful share of low-sensitivity internal traffic, the honest answer may be that you are overpaying for an exemption you do not need.

V. A Default Is a Configuration. A Prohibition Is a Covenant.

Four providers, four different legal instruments, one converging commercial posture, and far more variance in the strength of the promise than procurement generally assumes.

Anthropic's Commercial Terms of Service, effective June 17, 2025, contain the strongest language we located. Section B states flatly: "Anthropic may not train models on Customer Content from Services." The same section adds that "Anthropic disclaims any rights it receives to the Customer Content under these Terms" (Anthropic, 2025a). There is no default setting to be configured and no opt-in path described. The bar sits inside the agreement.

OpenAI's enterprise privacy documentation states: "By default, we do not use your business data for training our models. If you have explicitly opted in to share your data with us (for example, through our opt-in feedback mechanisms) to improve our services, then we may use the shared data to train our models." It further notes that "OpenAI may securely retain API inputs and outputs for up to 30 days to provide the services and to identify abuse," with zero data retention available

for eligible endpoints and qualifying use cases (OpenAI, 2026b). This is a strong and clearly stated position. It is also a default, and a default is a setting that someone in your organization can change.

Microsoft enumerates a set of negatives for models sold by Azure in Foundry: prompts, completions, embeddings, and training data "are NOT available to other customers," "are NOT available to OpenAI or other providers of Models sold by Azure," "are NOT used by providers of Models sold by Azure to improve their models or services," and "are NOT used to train any generative AI foundation models without your permission or instruction" (Microsoft, 2026). The qualifier in that last clause recurs throughout the document, and it is the same opt-in door OpenAI leaves open, phrased as a carve-out rather than as a default. Google Cloud's Service Specific Terms carry a Training Restriction to similar effect, committing that customer data will not be used to train or fine-tune models without prior permission or instruction (Google Cloud, n.d.-a). We paraphrase that clause rather than quoting it, because the operative sentence was not retrieved directly in this pass.

Set the four side by side and the distinction is sharp. A default is a configuration. A prohibition is a covenant. One can be changed by an administrator with the right permissions and a plausible reason. The other requires renegotiating a contract. Most vendor-assessment checklists we have seen ask whether the provider trains on customer data, record a "no," and move on, which treats the four instruments as equivalent when they are not.

The implication for leaders. This one is runnable on a Monday morning and costs an hour. Pull the operative clause, not the marketing page, for each model provider your organization uses. Classify each as covenant, default, or permissioned carve-out. For every default and every carve-out, identify by name the person who has authority to flip it and what approval that would require. If nobody can answer that last question, your risk register is describing a stronger promise than the one you actually hold.

VI. They Dropped the Exit Fee, and Most Enterprises Stayed Put

The cloud egress fee was, for roughly a decade, the most-cited technical lock-in of the platform era. It was quantifiable, it appeared on invoices, and it was the standard illustration of why data gravity kept workloads where they were. It is also the one platform-era switching cost that was actually abolished, which makes it the closest thing this field has to a controlled experiment.

Article 29 of Regulation (EU) 2023/2854, the Data Act, is titled "Gradual withdrawal of switching charges." From January 11, 2024, providers may impose only reduced switching charges not exceeding the costs directly incurred. From January 12, 2027, they may impose no switching charges at all for the switching process. Article 34(2) separately permits cost-based egress charges for ongoing parallel multi-cloud use, capped at the costs incurred; the prohibition targets the one-off exit (European Union, 2023). The market moved ahead of the deadline. Google Cloud announced free network data transfer for customers migrating off the platform on January 11, 2024, applying "to all customers globally" (Google Cloud, 2024), and AWS followed on March 5, 2024, waiving data-transfer-out charges "when you want to move outside of AWS" (Amazon Web Services, 2024). The relief in both cases is conditional in ways worth understanding before relying on it: both offers apply to customers actually leaving the platform, both are initiated through a request process, and

AWS's terms as updated in September 2025 give eligible customers ninety days to complete the migration (Amazon Web Services, 2024; Google Cloud, 2024).

The contractual cost of leaving a hyperscaler was then capped by statute and voluntarily zeroed by two of the three largest providers, more than two years before the statutory deadline required it. Enterprises still largely did not move. Multi-cloud remained a slide rather than a practice for most organizations, and the workloads that were going to stay stayed.

Remove the contractual cost, watch nobody leave, and you have learned that the cost was never contractual. It was the identity and access management model that had grown around one provider's primitives. It was the infrastructure-as-code that assumed one provider's resource names. It was the three years of operational scar tissue. It was the fact that the people who knew how to run the thing knew how to run it in one place.

The deeper parallel is enterprise resource planning, where the same lesson comes with a nine-figure price tag and an external auditor's signature. Birmingham City Council, the largest local authority in Europe, began replacing a 1997-vintage SAP system with Oracle Fusion in October 2019, on an expected budget of £19 million and an expected go-live in December 2020. The system went live in April 2022, the projected cost rose to £131 million including a subsequent re-implementation, and in January 2025 the auditors judged the council unlikely to have a fully functioning finance system until at least 2026 (Clark, 2025). An Audit Reform Lab analysis at the University of Sheffield, commissioned by three trade unions, put the wider loss at £216.5 million by April 2026 once anticipated savings already written into spending plans are counted, including £69 million of 2023/24 savings written off because budgets could not be monitored (Audit Reform Lab, 2024). From April 2022 until late summer 2023 the council had no audit trail in the accounting system and could not say whether financial fraud had taken place, and it has been unable to file auditable accounts since go-live (Clark, 2024, 2025). A Section 114 notice, the local-government equivalent of declaring insolvency, followed in September 2023, with the Oracle programme one contributing cause among several.

Grant Thornton, the council's local auditor, issued a report in the public interest on the ERP implementation on February 20, 2025, under Section 24 of the Local Audit and Accountability Act 2014 (Grant Thornton, 2025). Its findings are the ones that matter here, and not one of them is technical. The council authorized go-live with the design unresolved and with known deficiencies in the system's functional effectiveness. Governance carried little independent assurance, little senior-officer ownership, little oversight by elected members, and no strong design authority. The council did not equip itself with in-house skills in key programme-management roles, and it engaged the operational teams who would actually use the system too late in the business-change process. A leadership culture that did not want to hear bad news had significantly undermined risk management, and senior officers carried conflicting priorities to keep to budget, avoid delay, and protect reputations (Clark, 2025). Those findings are paraphrased from reporting on the audit rather than quoted from the report itself.

The U.S. Department of Veterans Affairs supplies the same lesson at national scale. VA signed with Cerner in May 2018 at a contract value of roughly \$10 billion, later revised above \$16 billion. The Institute for Defense Analyses estimated in 2022 that lifecycle costs for the electronic health record modernization would total \$49.8 billion, comprising \$32.7 billion for thirteen years of implementation and \$17.1 billion for fifteen years of sustainment. A later lifecycle estimate of

roughly \$37 billion was provided to congressional panels but not shared with the Government Accountability Office, despite an updated cost estimate standing as a GAO priority recommendation. Deployments were paused in 2023 over user concerns and resumed in 2026 after roughly a three-year halt (U.S. Government Accountability Office, 2026). The governance failure here has very little to do with vendor selection. Somewhere across those eight years, the organization lost the ability to price its own dependency.

One honest limit on this material. Published cost figures for health-record systems, which range from about \$3.88 million at Veterans Memorial Hospital to roughly \$800 million at the largest systems, are implementation costs rather than switching costs (Becker's Hospital Review, n.d.). We did not locate a peer-reviewed study isolating the counterfactual cost of moving from an installed system to a competitor's, and we are not going to imply one exists. The pattern across these cases is directional evidence, and it is consistent, but the specific switching number has not been measured by anyone.

The implication for leaders. When your team argues about model portability, ask what fraction of the estimated switching cost is contractual and what fraction is organizational. If nobody has separated the two, the estimate is not usable. The contractual portion tends to shrink on its own under regulatory and competitive pressure, on a timescale of about two years. The organizational portion only shrinks if someone deliberately shrinks it.

VII. Harvey, and What Bargaining Power Actually Looks Like

The strongest counter-case in this research is a named enterprise buyer describing exactly the risk calculus the continual-learning thesis assumes enterprises will surrender.

Harvey, the legal AI platform, published a first-party architectural statement in March 2026 explaining why it is multi-model by design. Its author, Niko Grupen, names "model provider risk" as a distinct enterprise risk category, alongside the operational, regulatory, and geopolitical disruptions that produce it. The architectural rationale is redundancy: "Harvey's multi-model architecture provides structural redundancy. If one provider experiences capacity constraints, service degradation, or an outage, Harvey can route work to an alternative model without disrupting the user's workflow" (Grupen, 2026).

Harvey routes by task rather than by preference. Opus 4.6 handles agentic reasoning and deep multi-step legal analysis. GPT-5.2 handles capability-aware, sourced long-form output. Sonnet 4.6 and Gemini 3 Flash handle latency-sensitive, high-volume surfaces such as Vault. The firm also manages regional availability asymmetry, noting that Opus 4.6 was available in Australia where certain other frontier models were not, and it delegates model choice down to workspace administrators and to individual practitioners.

Then comes the sentence that decides the argument of this paper: "Harvey requires all model providers to meet the same non-negotiable standards: zero data retention, no human review of customer data, and no training on customer inputs. Our multi-provider relationships give us leverage to enforce these commitments consistently, and the flexibility to respond if a provider's posture changes" (Grupen, 2026).

A buyer with real multi-vendor optionality uses that optionality to forbid the training the moat requires. That is the mechanism by which the moat fails to form, and the claim deserves stating precisely, because it differs from the usual one. Plenty of enterprises carry heavy workload-level switching costs, and Sections III and VI locate where they sit. What matters is that the buyers sophisticated enough to be worth locking in are the same buyers with the bargaining power to refuse, and that refusing is cheap for them, because multi-provider routing is something they were building anyway for reliability reasons.

The corollary is uncomfortable for everyone else. Harvey's position is available to any organization willing to build a second integration and maintain a provider-neutral evaluation suite. Most have not, and their exposure traces back to an engineering decision made eighteen months ago by someone who reasonably wanted to ship, rather than to any lab's ambition.

The implication for leaders. Multi-provider routing is a governance instrument before it is an availability instrument. The organizations that can enforce data terms across providers are the ones that already run against more than one. If your architecture cannot route a workload to a second provider without a project, your data terms are whatever your provider decides they are at renewal.

What the Record Adds Up To

Four things are true at once, and the discussion tends to hold only one of them at a time.

The switching cost enterprises actually carry is real, and it is organizational. It binds at the level of the validated workload rather than the provider relationship: new spend reallocates when capability moves, while the deployed workload carries the evaluation suites, cleared reviews, and tuned scaffolding that make moving expensive. That cost exists without any weight-level adaptation and would survive the disappearance of every proprietary model on the market.

The technical mechanism the continual-learning thesis depends on is receding rather than advancing in the commercial market, on published schedules, at OpenAI, and the research literature is candid that the underlying problem is unsolved.

The data-for-capability bargain exists today, priced and in writing, aimed at consumers and free-tier developers, with enterprises excluded by the sellers and paying for exemption.

And the historical record on contractual switching costs, across cloud egress, ERP, and health records, is that the contractual portion is the portion that gets abolished, and abolishing it changes very little.

The honest statement of the counter-position deserves space, because all three of the blocks described in this paper are contingent. Contracts get renegotiated when a capability differential grows large enough to justify it. Cloud intermediaries could offer per-tenant continual adaptation themselves and capture the advantage at the hyperscaler level rather than the lab level, which would be a more concentrated outcome rather than a less concentrated one. OpenAI's enterprise exclusion is plausibly a compliance-posture decision that reverses the day a competitor makes the offer credible. None of that is refuted by the current record. The defensible claim is narrower than "this will not happen." It is that this is not happening yet, the observable trend is the opposite, and

an enterprise decision that trades away training rights today for a capability advantage that does not yet exist in any purchasable form is a decision made on a forecast.

One counter-argument deserves to be named and then set aside, because it is the one most commonly reached for and the measurement evidence does not support it. Open-weight models are frequently offered as the reason none of this matters, on the theory that commoditization will dissolve any lab's pricing power. In the long run that ceiling is probably real. In this window it is not operating. Stanford's 2026 AI Index reports the gap between the top closed and top open models reopening after it nearly closed in 2024: 0.5 percent in August 2024, 3.3 percent as of March 2026 (Stanford HAI, 2026). Menlo reports enterprise open-source model share falling from 19 percent to 11 percent between its 2024 and 2025 surveys (Menlo Ventures, 2025). A strategy that assumes open weights as the escape hatch is assuming something the two best measurement sources currently contradict.

What to Do

Five things, none of which requires a new framework, a consulting engagement, or a position on when continual learning arrives.

Separate the two halves of your switching cost and put a number on each. The contractual half is the exit fees, the license terms, the minimum commitments, and the data-extraction charges. The organizational half is the evaluation suite, the prompt and tool scaffolding, the retrieval index, the cleared security review, the signed agreements, and the people. Do this once, at the workload level rather than the enterprise level, and the estimate will be uncomfortable in a useful way. If the contractual half is more than a quarter of the total, either your estimate is wrong or you have an unusual contract.

Classify each provider's training clause as a covenant, a default, or a permissioned carve-out, and name the person who can change each default. Section V describes the exercise. It takes an hour and produces a risk-register entry that is actually true.

Decide the training-rights question per workload rather than per enterprise. A blanket no-training posture across every workload is expensive and, for a meaningful share of internal traffic, unnecessary. A blanket yes is indefensible. The decision belongs at the level of the data class, and somebody should own writing that policy down.

Keep one workload live on a second provider at all times, and keep your evaluation suite provider-neutral. Treat it as a standing capability rather than a migration plan: priced as insurance, exercised quarterly. This is the concrete form portability discipline takes, and Harvey's position shows what it buys, which is the ability to enforce terms rather than accept them.

Put your providers' deprecation calendar next to your model-governance calendar and look at where they collide. The October 23, 2026 and November 30, 2026 shutdown dates on OpenAI's deprecations page are the near-term test of whether your organization can revalidate a replacement model inside a provider's timetable. That capability is the same capability portability requires, which makes this the rare governance investment that pays twice.

The Next Twenty-Four Months

Six predictions, stated so they can be checked rather than admired.

No major provider will offer continual updating of a customer-specific base model from enterprise production traffic as a generally available, contracted product before August 2028. Falsified by a general-availability announcement with published terms, not by a research preview.

OpenAI's January 6, 2027 fine-tuning cutoff will hold, and no comparable self-serve base-weight tuning product will replace it at that provider inside the window. Falsified by a reversal or an equivalent replacement offering.

The EU Data Act's zero-switching-charge obligation takes effect on January 12, 2027, and model-provider switching charges will not become a live enterprise complaint when it does, because there are very few to abolish. The complaints will be about evaluation continuity and deprecation schedules instead.

At least one provider will test an enterprise-facing data-for-price offer inside the window. The takers will be mid-market firms without negotiated data-processing agreements rather than large regulated enterprises. Falsified by a named Fortune 500 firm publicly accepting such terms.

Model deprecation, rather than model lock-in, will be the first of these issues to reach an audit committee in a large regulated enterprise. Falsified by the reverse ordering showing up in public disclosures or enforcement actions.

Open-weight models will not close the enterprise gap inside this window. Falsified by the AI Index gap narrowing below one percent or enterprise open-source share returning above 19 percent in the next Menlo survey.

No one has run this experiment at scale yet, and anyone who claims to know how the continual-learning question resolves is guessing at a timeline the research literature has not committed to. The direction of the practices that keep an organization free to act is already clear enough to act on. Every one of them, a provider-neutral evaluation suite, a retrieval index you can carry, a second integration kept warm, a data-class policy on training rights, an operative-clause inventory with names attached, pays off whether or not the moat ever forms. That is the useful property of organizational capability, and it is why *The Transformation Myth* argued for moving boldly rather than waiting for the picture to resolve (Kane et al., 2021). Build the capability because it describes an organization that can still choose, whatever the labs ship next.

The exit fee is gone from the cloud contracts. What kept the workloads in place is still there, and everybody could have told you it would be.

Methodology Note

This paper is built on primary documents rather than on secondary commentary. Its evidence base consists of eight provider contract and policy documents read directly (Anthropic's commercial and consumer terms, OpenAI's enterprise privacy page, deprecations page, and data-sharing programme documentation, Google's Gemini API Additional Terms, Microsoft's Foundry data-privacy documentation, and AWS's Bedrock data-protection guide), two hyperscaler egress

announcements read directly, one statutory instrument, one peer-reviewed NeurIPS paper and two research sources on continual learning, one operator disclosure of serving economics, three research-house measurement reports, one government audit series, and one first-party enterprise buyer statement. Four limits should be stated plainly. First, the Grant Thornton report in the public interest on Birmingham City Council's ERP implementation was located and cited at its published address on the council's own site, but its findings are summarized here from contemporaneous reporting rather than from the document's own text, so no wording from it is presented as a quotation; the same care applies to Google Cloud's Training Restriction clause, which is described rather than quoted. Second, the Menlo Ventures surveys are vendor-adjacent venture research based on self-reported responses (495 U.S. decision-makers in 2025; 600 in 2024), and Menlo is an investor in Anthropic, whose share gains its surveys report, so the figures are used here for market structure rather than precision; note also that the \$11.5 billion 2024 spend figure is Menlo's later inference-excluded restatement, not the figure its 2024 report originally published. Third, DeepSeek's serving disclosure is self-reported and unaudited, and its theoretical margin figure is not treated here as an achieved result. Fourth, a concurrency threshold widely quoted in commentary on sparse-model serving economics could not be traced to any public primary source and is therefore characterized in Section I as unestablished rather than repeated. The triggering essay is treated throughout as the prompt for this research rather than as evidence within it.

References

- Amazon Web Services. (n.d.). *Data protection in Amazon Bedrock*. Amazon Bedrock User Guide. <https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html>
- Amazon Web Services. (2024, March 5). *Free data transfer out to internet when moving out of AWS* (S. Stormacq). AWS News Blog. <https://aws.amazon.com/blogs/aws/free-data-transfer-out-to-internet-when-moving-out-of-aws/>
- Anthropic. (2025a). *Commercial terms of service* (effective June 17, 2025). <https://www.anthropic.com/legal/commercial-terms>
- Anthropic. (2025b, August 28). *Updates to consumer terms and privacy policy*. <https://www.anthropic.com/news/updates-to-our-consumer-terms>
- Becker's Hospital Review. (n.d.). *From \$3.8M to \$800M: The cost of EHR implementations at 12 hospitals*. <https://www.beckershospitalreview.com/healthcare-information-technology/ehrs/from-15m-to-800m-the-cost-of-ehr-implementations-at-3-hospitals/>
- Audit Reform Lab. (2024, August). *Value for money and accountability: Birmingham City Council*. University of Sheffield Management School. <https://auditreformlab.group.shef.ac.uk/downloads/bcc-report-4-08-24.pdf>
- Clark, L. (2024, August 20). *City council faces £216.5M loss over Oracle system debacle*. The Register. <https://www.theregister.com/software/2024/08/20/birmingham-council-faces-2165m-loss-over-oracle-debacle/1519213>
- Clark, L. (2025, January 29). *Oracle finance system at Europe's largest city council still falls short 2.5 years later*. The Register. <https://www.theregister.com/software/2025/01/29/birminghams-oracle-finance-system-falls-short-25-years-on/548607>
- DeepSeek. (2025, March 1). *Day 6: One more thing, DeepSeek-V3/R1 inference system overview*. Open Infra Index. https://github.com/deepseek-ai/open-infra-index/blob/main/202502OpenSourceWeek/day_6_one_more_thing_deepseekV3R1_inference_system_overview.md

- European Union. (2023). *Regulation (EU) 2023/2854 of the European Parliament and of the Council (Data Act), Article 29: Gradual withdrawal of switching charges*. <https://eur-lex.europa.eu/eli/reg/2023/2854/oj>
- Google. (2026). *Gemini API additional terms of service* (effective March 23, 2026). <https://ai.google.dev/gemini-api/terms>
- Google Cloud. (2024, January 11). *Cloud switching just got easier: Removing data transfer fees when moving off Google Cloud* (A. Zavery). Google Cloud Blog. <https://cloud.google.com/blog/products/networking/eliminating-data-transfer-fees-when-migrating-off-google-cloud>
- Google Cloud. (n.d.-a). *Service specific terms*. <https://cloud.google.com/terms/service-terms>
- Google Cloud. (n.d.-b). *Tune Gemini models by using supervised fine-tuning*. Vertex AI documentation. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-use-supervised-tuning>
- Google Research. (2025, November). *Introducing Nested Learning: A new ML paradigm for continual learning*. <https://research.google/blog/introducing-nested-learning-a-new-ml-paradigm-for-continual-learning/>
- Grant Thornton UK LLP. (2025, February 20). *Report in the public interest: Birmingham City Council ERP implementation*. Birmingham City Council. https://www.birmingham.gov.uk/download/downloads/id/30136/public_interest_report_-_erp_implementation_2025.pdf
- Grupen, N. (2026, March 10). *Why Harvey is multi-model by design*. Harvey. <https://www.harvey.ai/blog/why-harvey-is-multi-model-by-design>
- Kane, G. C., Nguyen Phillips, A., Copulsky, J. R., & Andrus, G. R. (2019). *The technology fallacy: How people are the real key to digital transformation*. MIT Press.
- Kane, G. C., Nanda, R., Nguyen Phillips, A., & Copulsky, J. R. (2021). *The transformation myth: Leading your organization through uncertain times*. MIT Press.
- Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2025). *An empirical study of catastrophic forgetting in large language models during continual fine-tuning* (arXiv:2308.08747). <https://arxiv.org/abs/2308.08747>
- Menlo Ventures. (2024, November 20). *2024: The state of generative AI in the enterprise*. <https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/>
- Menlo Ventures. (2025, December 9). *2025: The state of generative AI in the enterprise*. <https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/>
- Microsoft. (2026). *Data, privacy, and security for models sold by Azure in Microsoft Foundry*. Microsoft Learn. <https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/openai/data-privacy>
- OpenAI. (2026a). *Deprecations*. OpenAI API documentation. <https://developers.openai.com/api/docs/deprecations>
- OpenAI. (2026b). *Enterprise privacy at OpenAI*. <https://openai.com/enterprise-privacy/>
- OpenAI. (2026c). *Sharing feedback, evaluation and fine-tuning data, and API inputs and outputs with OpenAI*. OpenAI Help Center. <https://help.openai.com/en/articles/10306912-sharing-feedback-evaluation-and-fine-tuning-data-and-api-inputs-and-outputs-with-openai>
- Patel, D. (2026, August 7). *8 predictions for the era of continual learning*. Dwarkesh Podcast. <https://www.dwarkesh.com/>
- Stanford HAI. (2026). *The 2026 AI Index report*. Stanford Institute for Human-Centered Artificial Intelligence. <https://hai.stanford.edu/ai-index/2026-ai-index-report>
- U.S. Government Accountability Office. (2026). *VA electronic health record modernization: Critical actions needed to support accelerated system deployments* (GAO-26-108812). <https://files.gao.gov/reports/GAO-26-108812/index.html>

Zweiger, A., Pari, J., Guo, H., & Kim, Y. (2025). *Self-adapting language models* (SEAL). Advances in Neural Information Processing Systems 38. https://papers.neurips.cc/paper_files/paper/2025/file/6b41e04c41726e2a60e456d0a2b961ab-Paper-Conference.pdf

Copyright, reuse, and citation

© 2026 Gerald C. Kane. This paper is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), <https://creativecommons.org/licenses/by-nc/4.0/>.

You may share and adapt it for noncommercial purposes, provided you credit Gerald C. Kane as the author, link to the license, and indicate any changes you made. You may not present this material, or adaptations of it, as your own work, and you may not sell it or use it in paid training or consulting without written permission.

Cite as: Kane, G. C. (2026). *They Dropped the Exit Fee*. <https://www.profkane.com/papers/they-dropped-the-exit-fee.pdf>

Permissions and commercial use: jerry@profkane.com