Online ISSN: 3104-9338
Print ISSN: 3104-932X

From Corpus to Model: Governing the Knowledge Supply Chain of Large Language Models

Ruyue Yang, Xinyu Song*
Geely University of China, ChengDu, China, 641423

Received: August 22 2026

Revised: August 30, 2026

Accepted: August 30, 2026

Published online: August 30, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 9 (September 2026)

* Corresponding Author: Xinyu Song (songxinyu@guc.edu.cn)

Abstract. Large language models acquire most of their knowledge from sources—web corpora, licensed archives, curated datasets—whose governance traditionally depends on a capability that parametric learning destroys: the ability to take knowledge back. This survey reviews 18 core sources (2018–2024) through an original knowledge supply chain framework that traces knowledge from upstream sourcing and licensing, through midstream injection into model parameters or retrieval indexes, to downstream use and recall, with a governance control point attached to each stage; reported findings are quoted only from primary sources and court records. Three findings emerge. Upstream, the data foundation is weakly documented: large-scale audits find that most open datasets do not convey their own licensing terms, and web-scale corpora contain removed and toxic content at unrecorded rates. Midstream, the four knowledge-injection channels differ enormously in reversibility—retrieval is fully reversible, model editing partially so, pre-training effectively not at all—yet systems treat them interchangeably. Downstream, removal is the weakest control: machine unlearning remains unverified at scale exactly as litigation begins to demand it. We map documentation standards, reversibility grading, and verified deletion onto the chain's stages as actionable control points.

Keywords: *data provenance; machine unlearning; retrieval-augmented generation; licensing; dataset documentation; copyright*

1. Introduction

1.1. The Lawsuit That Exposed a Missing Capability

On December 27, 2023, The New York Times Company sued Microsoft and OpenAI in the

Southern District of New York, alleging that millions of its articles were used to train large language models without permission and that the models can reproduce near-verbatim stretches of its journalism [The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y.)] [1]. Whatever its outcome, the complaint exposes a structural asymmetry that no verdict can repair. In ordinary data governance, inclusion is reversible: a record entered without authorization can be deleted, and deletion propagates to every system that respects the source of truth. In a trained language model, the copy that matters is not a record but a distributed change in billions of parameters, entangled with everything else the model learned from the same corpus. There is no query that removes it, no index to drop it from, and no rollback to a pre-training state that would not also erase everything else. Knowledge, once injected into parameters, has left every regime in which "remove it" is a meaningful instruction.

The procedural posture of the case amplifies the asymmetry rather than softening it. The complaint's requested relief reportedly includes destruction of models trained on infringing copies—an equitable remedy that, in ordinary product governance, corresponds to a full recall of a manufactured good, an event rare and catastrophic enough that entire insurance and inspection industries exist to prevent it [1]. Whatever disposition the court reaches, the industry's behavior since filing suggests it has accepted the underlying diagnosis: licensing negotiations between model developers and content owners have proliferated, opt-out registries have become negotiating instruments, and disclosure of training-data composition has moved from an academic demand to a term of commerce. These are upstream and contractual responses to a defect the industry has not learned to remediate once knowledge is already in the weights—which is precisely why the case matters beyond its parties: it documents, in the most authoritative forum available, that the governance toolkit currently ends where training begins.

This survey treats that asymmetry as the defining governance problem of large language models and organizes the field's evidence around it. The organizing image is a supply chain: knowledge is sourced upstream, injected midstream, used downstream, and—if governance requires—recalled, each stage with its own documented failure modes and its own candidate control points.

1.2. Scope and Method

We focus on the governance of knowledge in large language models—where it comes from, how it enters, how it stays correct, how it leaves—rather than on the deployment governance of model applications (bias in hiring, content moderation), which is treated extensively elsewhere. We searched arXiv and major machine-learning venues for 2018–2024 primary

sources, complemented by court records and European Union legislative texts. Sources were retained if they (i) audit or document the provenance and licensing of training data, (ii) propose or evaluate a mechanism that injects, corrects, or removes knowledge, or (iii) constitute primary governance instruments. Purely positional commentary on AI copyright was excluded; the survey quotes the litigation only for its documented factual claims. This procedure yielded the 18 core sources organized below.

The verification protocol deserves a sentence of its own, because governance writing inherits the defects it audits unless it disciplines itself. Every empirical claim in what follows is quoted from the audited source itself—benchmark papers for benchmark numbers, audits for audit statistics, court records for litigation facts—rather than from commentary about them; each source's author list, title, and venue were checked against its primary record before inclusion; and where a citation's provenance could not be verified with confidence, the source was dropped rather than cited on recollection. Two consequences follow. First, the survey's evidential base is deliberately narrower than the field's literature—it contains no secondary summaries of failure rates, no numbers remembered from other surveys—and its figures can be traced to a primary document in one step. Second, and more usefully for the reader, the verification burden itself models the survey's subject: the demand that a claim carry its provenance is exactly the upstream record this survey will argue the model-knowledge supply chain lacks. The survey asks a standard of the field's knowledge claims that it attempts to meet for its own. The survey deliberately overlaps with, but is distinct from, agent-memory surveys: memory systems govern what an agent writes down between sessions, while this survey governs what a model is—its parameters, corpora, and retrieval sources.

1.3. Positioning and the Supply Chain Framework

Three prior syntheses each illuminate one region of the problem without spanning it. Bommasani et al. introduced the foundation-model frame and catalogued the societal stakes of models trained on undifferentiated data, but as a broad agenda document it does not organize technical evidence by governance stage [2]. Ji et al. survey hallucination in natural language generation, giving the field its standard intrinsic/extrinsic taxonomy of unfaithful output—a midstream-to-downstream maintenance failure treated in isolation from where the knowledge came from [3]. The Data Provenance Initiative audits dataset licensing and attribution at scale—an upstream audit that stops at the dataset boundary and says nothing about what training does with what it documents [4]. Table 1 contrasts the three with this survey.

Table 1. Coverage comparison between prior syntheses and this work.

Dimension	Bommasani et al. [2]	Ji et al. [3]	Data Provenance Initiative [4]	This survey
Primary scope	Foundation-model agenda	Hallucination taxonomy	Dataset licensing audit	Knowledge lifecycle governance
Chain coverage	Broad, non-staged	Midstream–downstream	Upstream only	Upstream through recall
Failure evidence	Framing arguments	Taxonomy + mitigations	License audit statistics	Quantified, all stages
Governance output	Research agenda	Mitigation methods	Compliance tooling	Control points per stage

The framework's central term deserves definition before it is used. A governance control point is a stage-attached intervention at which three properties can be established: attribution (the intervention names the knowledge it governs—this corpus, this channel, this output), verification (some procedure can check that the intervention did what it claims—audit, probing, record inspection), and enforceability (an actor with authority—court, regulator, procurer, operator—can compel or reward compliance). Mature supply-chain regulation, from food safety to electronics, is built from exactly such points: documented inputs, graded intermediates, traceable outputs, and recall authority at each stage. Read against that template, the model-knowledge chain fails not at any single point but in a specific pattern: attribution is weakest upstream (untraceable corpora), verification is weakest at recall (unproven unlearning), and enforceability, which exists in law, currently attaches to nothing it can grip. The survey's analytical claim is that these are engineering specifications, not philosophical puzzles—each section will name what a working control point at its stage would measure.

The supply chain framework gives each stage a control question. Upstream: can the provenance and permission of every source be established? Midstream: through which channel does knowledge enter, and how reversible is that channel? Downstream: does output remain faithful to sources that may themselves have changed? Recall: can knowledge be removed when law or correction demands it? Section 2 takes the stages in order; the framework's payoff is that remedies attach to stages—a datasheet cannot fix a ripple effect, and an unlearning benchmark cannot fix an unlicensed corpus. Figure 1 previews the chain, its documented failures, and its control points.

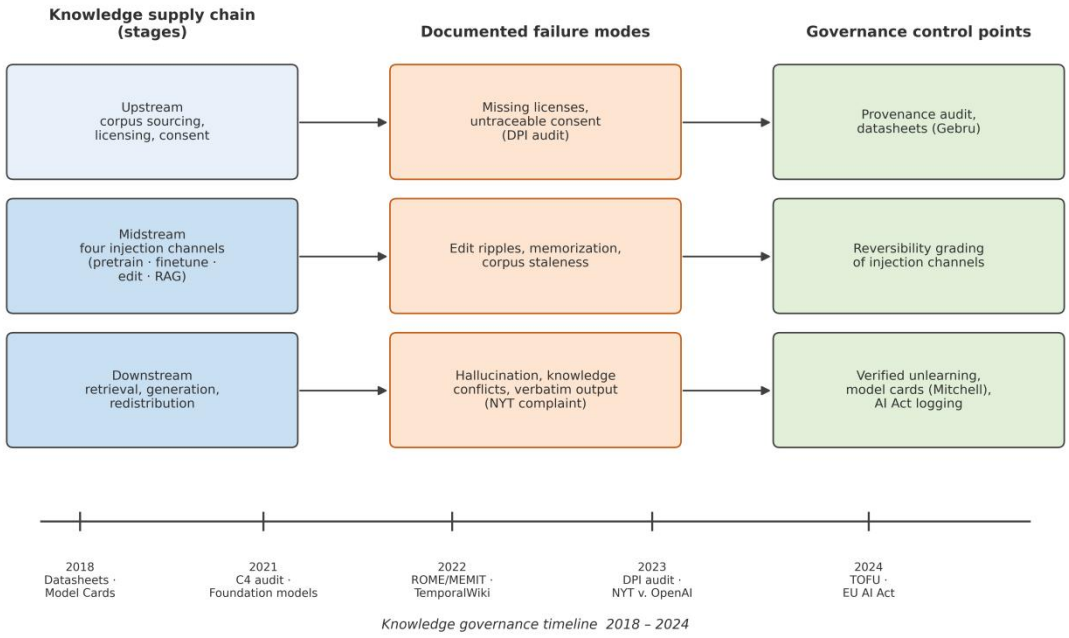

Figure 1. The knowledge supply chain of large language models. Left: the three lifecycle stages plus recall. Middle: documented failure modes at each stage. Right: governance control points. Bottom: the field's timeline, 2018–2024.

2. The Upstream: Provenance Without Permission

The chain begins with corpora that are, by the field's own audits, weakly governed. Dodge et al. supplied the first systematic documentation of C4, the Colossal Clean Crawled Corpus behind many prominent models, and found a dataset assembled by heuristic filtering rules with essentially no provenance discipline: pages disappear after scraping at unrecorded rates, filtering silently drops whole registers of text, and the surviving content skews toward particular dialects and publishers in ways no one had measured [5]. The documentation gap is not cosmetic; it means that when a model trained on such a corpus later emits a copyrighted or toxic passage, the data to trace where that passage came from does not exist in any usable form.

The licensing picture is weaker still. The Data Provenance Initiative audited nearly 2,000 widely used open datasets and their thousands of source documents, and found that more than half of them fail to convey license information usable in practice—terms are missing, misdescribed, or attached to sources they do not actually govern [4]. In supply-chain terms, most of the field's raw material arrives without a paper trail: not necessarily acquired wrongfully, but unauditable, which for governance purposes is the operative defect. An unauditable input cannot be excluded ex ante, priced, or recalled ex post; it can only be litigated after the fact, which is precisely the position the NYT complaint [1] occupies. The upstream control point is therefore documentary before it is legal—establishing what entered, under what terms—

because every downstream control presupposes that record.

The texture of the C4 audit is worth conveying, because it shows that the documentation failure is systematic rather than a matter of a few negligent sources. Dodge et al.'s methodology was itself an act of reconstruction: they manually inspected samples of the corpus and traced documents back to their URLs, classifying what remains of each origin page. A large majority of the traced documents no longer resolve to their original content—the pages were deleted, moved, or changed after crawling—so the corpus now contains text whose source context is unrecoverable from the live web, and the audit's classification of content type and quality had to proceed from the scraped snapshots alone [5]. Two findings carry particular governance weight. Filtering, the audit shows, was where most silent loss occurred: the corpus's cleaning heuristics, designed to remove boilerplate and gibberish, differentially removed entire registers of text, with the surviving document distribution skewed toward a narrow band of publishers and registers—and the corpus's documented racial, gender, and toxicity profiles follow that skew rather than any property of the web itself [5]. In supply-chain terms, the cleaning step was a processing operation that changed the product's composition without leaving a record of what it removed: the model trained on C4 learned from a filtered world while the field's documentation described an unfiltered one.

The Data Provenance Initiative's licensing audit has the same reconstructive character at larger scale. Auditing nearly 2,000 open text datasets—more than 24,000 source documents in total across them—the initiative found that license information is most often absent entirely for crawled sources, that licenses are at times attached to collections but not to the individual sources whose terms they should govern, and that misdescription runs in both directions: sources described as more permissive than they are, and public-domain material unnecessarily restricted by dataset-level terms [4]. The audit's own framing is the supply-chain one—its authors describe the result as a transparency crisis in the "data supply chain" of AI—and its statistics give the upstream control question of Section 1.3 its empirical content: attribution fails at the majority of sources, verification fails wherever terms are misdescribed, and enforceability has nothing to attach to in either case [4].

It is worth being precise about why the audit came after the fact rather than before. The incentives of corpus construction point the wrong way: scraping is cheap, comprehensive, and fast, while permission tracking is slow, partial, and costly—and the training run that consumes the corpus can begin the day the crawl finishes. Auditing, by contrast, is publishable exactly when it is damning, so the institutions best placed to document their inputs have the least reason

to. This is the same asymmetry the validation literature documented in clinical AI, reproduced one stage earlier in the chain: the defect is not that documentation is technically hard, but that its costs fall on the party that benefits from its absence. Governance instruments—mandatory disclosure tiers, procurement requirements that condition acquisition on documentation—exist in mature supply chains precisely to override that incentive, and the upstream record shows what their absence has produced.

The current partial correction—the spread of opt-out signals and licensing deals—has the structure of a bargaining process conducted under the asymmetry the audits document. A content owner can exclude future crawls, but has no practical way to know whether past inclusion occurred, no inventory of what was taken, and therefore no basis for pricing what was lost; a model developer can offer a license for future use, which costs little against the alternative of litigation over the past. The NYT complaint crystallizes the endgame of that imbalance: when voluntary signals failed to produce accounting, the recourse is a demand, backed by court process, for precisely the inventory the supply chain never kept [1]. Read this way, the litigation is not an anomaly interrupting an otherwise functioning market; it is what markets do when the goods are traded without documentation and the audit arrives after consumption. The upstream lesson for governance is correspondingly blunt: documentary control points must be installed before the bargaining, not litigated after it, because every year of undocumented training increases the stock of knowledge whose provenance can never be established retroactively [4].

3. The Midstream: Four Injection Channels of Unequal Reversibility

3.1. Retrieval: The Fully Reversible Channel

Retrieval-augmented generation keeps knowledge outside the parameters entirely: documents live in an index, the model consults them at query time, and generation is grounded in what was retrieved [6]. Governance treats this channel as familiar territory. Correcting the index corrects the model's knowledge instantly and completely; removing a document removes its influence; licensing attaches to identifiable items. The channel's governance problems—index quality, retrieval faithfulness—are ordinary information-management problems. This is why staleness and copyright pressure in practice push deployments toward retrieval: it is the only channel where the traditional toolkit still works as designed.

The channel's origin paper is worth rereading as a governance document, because its design

choices are governance choices. Lewis et al. demonstrated RAG on knowledge-intensive tasks—open-domain question answering among them—showing that a generator conditioned on documents retrieved from a Wikipedia index reached competitive accuracy while the knowledge itself remained a separate, swappable component [6]. The architectural separation is what the governance analysis prizes: knowledge lives in an enumerated index whose contents, versions, and access terms can be stated; updates are index operations with predictable scope; and every generated claim has an identifiable evidentiary source that preceded it. None of this makes retrieval a complete control point—Section 3.1's boundary statement already limited the claim to evidence rather than use—but it supplies the precondition every other channel lacks: the object of governance exists as a list. When a regulator, court, or operator asks "what does the system know and where did it come from?", the retrieval channel is the only one of the four that can answer at all [6].

The channel's governance boundary deserves an honest statement, because retrievability is not faithfulness. Grounding generation in a retrieved source does not guarantee the source is represented accurately—the model can mischaracterize what it retrieved—and the index inherits whatever defects its corpus carries, including the upstream documentation failures of Section 2. What retrieval guarantees is narrower and still decisive: the evidence in the system is identifiable, correctable, and removable as a matter of record-keeping, even where the model's use of that evidence remains imperfect. In supply-chain terms, retrieval relocates the hard problem from unmanageable (distributed parameter change) to merely difficult (retrieval and generation quality)—and relocating a problem to where existing governance applies is exactly what supply-chain regulation has always done.

3.2. Model Editing: The Partially Reversible Channel

Model editing techniques attempt to retrofit reversibility onto parameters. ROME located the weight positions mediating a factual association and rewrote them with a rank-one update [7]; MEMIT scaled the approach to thousands of associations at once [8]. As governance instruments, however, the editing family carries a documented defect: edits do not propagate. On the MQuAKE benchmark, models whose underlying fact was corrected still answer multi-hop questions with the pre-edit answer, because the correction never travels along the chains of inference that the fact supports [9]. Knowledge-conflict studies complete the picture: when edited or retrieved content contradicts parametric knowledge, models behave inconsistently, sometimes deferring to stale parameters against explicit evidence [10]. Editing, in short, is reversible in the small and unreliable in the large—a channel where "correction" is locally

verified and globally unverified.

The methodological lineage explains why the verification gap took the field by surprise. ROME's contribution was diagnostic as much as corrective: using causal-mediation analysis—corrupting inputs and tracing which internal activations carry the factual recall—it located the specific mid-layer feed-forward positions where a subject–relation–object association is stored, and demonstrated that a rank-one weight update there could implant or replace a single fact [7]. The verification the method offers is rigorous but pointwise: after the edit, the target probe ("The Eiffel Tower is located in") returns the new answer, and neighboring associations are spot-checked for collateral damage. MEMIT generalized the operation, distributing thousands of edits across the layers the analysis identified, which is what moved editing from a laboratory curiosity toward an operation one could imagine running against a deployed model [8]. Both papers verified exactly what their framings suggested—local insertion and local side effects—and both are honest about the scope. The governance failure came later, when the operation's pointwise verification was quietly read as a system-level guarantee: a patch whose regression tests pass at the patch site is not thereby safe in a dependency graph, and facts are nothing but dependency graphs. MQuAKE [9] supplied the missing integration test, and the field learned what software engineering has known since the first regression: local unit tests do not certify global behavior [7-9].

The governance imagination that editing research invites is worth naming, because the field repeatedly reaches for it: the edit as patch, a correction issued against a deployed model the way a security fix is issued against software. MEMIT's scaling result is what makes the patch metaphor plausible—thousands of updates, issued at once [8]. What MQuAKE's result is what makes it premature: a patch that does not propagate through the dependencies of what it changes is not a patch in any sense software engineering would recognize, and knowledge is almost nothing but dependencies—a corrected founding date implies corrected ages, durations, and sequences that no current editing method traces [9]. Until edit operations carry their dependency structure, the patch metaphor should be treated as aspiration rather than capability, and corrections that matter legally should be routed through channels whose propagation is guaranteed—that is, back through the index or the corpus.

3.3. Pre-training and Fine-tuning: The Effectively Irreversible Channel

What pre-training injects, nothing yet reliably extracts. Carlini et al. demonstrated that language models memorize training data deeply enough to regurgitate it—verbatim, including personally identifiable records—under black-box queries [11], and Brown et al. sharpened the

governance meaning of that fact: memorization is not itself the harm, but memorized content whose collection context promised privacy constitutes a standing disclosure waiting for the right query [12]. The channel's reversibility is effectively nil at pre-training scale; fine-tuning sits between channels, injectable knowledge that unlearning research targets but, as Section 5 shows, cannot yet verifiably remove. Fine-tuning's governance position is also the least examined: it is the channel through which organizations install their private, often regulated, knowledge into models, yet it carries none of the documentation tradition that upstream datasets have begun to acquire—an enterprise fine-tune is frequently the least documented acquisition of knowledge in the entire chain. Table 2 grades the channels.

The memorization evidence also quantifies the channel's exposure in a way that matters for governance planning. Carlini et al.'s extraction methodology—queries designed to elicit high-perplexity continuations, which surface memorized training text against the model's fluency prior—recovered verbatim personally identifiable records from a production-scale model, and the study's two structural findings are the ones a data controller must plan around: memorization per example increases with model scale, so the exposure grows with every capability generation; and repetition of an example in training drives memorization steeply, which makes duplication of sensitive records a controllable risk factor—the rare upstream decision that measurably changes the downstream harm [11]. Brown et al. then supplied the definitional correction that turns these results into a governance standard rather than a curiosity: privacy, they argue, cannot be evaluated in the abstract as "is this string in the weights," but only relative to the context in which the data was collected—a public court record and a private medical record may both sit in the weights, but only one was collected under a promise of confidentiality, and only the latter constitutes disclosure when extractable [12]. The standard relocates the compliance question from corpus statistics to collection context—which is to say, once more, to upstream documentation. A controller who cannot state under what expectations each corpus was gathered cannot say whether its model's memorization constitutes a standing breach; the midstream defect is inheritable from the upstream one [11,12].

Table 2. The four knowledge-injection channels graded by governance properties.

Channel	Knowledge location	Reversibility	Verified correction	Recall instrument
Retrieval [6]	External index	Full	Immediate, complete	Drop document
Model editing [7,8]	Weights (localized)	Partial	Local only; ripples fail [9]	Un-edit (unproven)
Fine-tuning	Weights (distributed)	Low	Unlearning, fragile (Section 5)	Unlearning
Pre-training [11]	Weights (entangled)	Effectively none	None demonstrated	Retrain (prohibitive)

The grading exposes the midstream's governance inversion: the easiest channel to inject at scale—pre-training—is the hardest to govern, while the governable channel—retrieval—is the one treated as an auxiliary. Deployment architecture is quietly a governance decision, and it is rarely framed as one.

The inversion has a specific casualty worth naming: the fine-tuning row of Table 2. Fine-tuning is the channel through which enterprises install their own regulated knowledge—customer records, internal policies, proprietary research, in many cases personal data subject to the obligations of Section 5—into a model, and it inherits pre-training's distributed irreversibility in miniature: the knowledge is in the weights, entangled with the base model's, removable only by unlearning of the fragile kind Section 5 measures. Yet the governance discussion concentrates on pre-training corpora and on retrieval indexes, while the enterprise fine-tune—the one injection most directly under a single controller's authority, and therefore the one where documentary discipline is actually achievable—proceeds with no datasheet, no channel statement, and no deletion story. A controller who cannot govern anything else can still govern what it fine-tunes on; that this channel is the least documented in practice is not a technical constraint but a choice the governance instruments have not yet forced anyone to revisit.

4. The Maintenance Problem: Staleness, Conflict, Hallucination

Injected knowledge ages, and the aging is measurable. TemporalWiki constructs a lifelong benchmark from consecutive Wikipedia snapshots and shows that models trained at one moment systematically fail on facts established or changed after it, with continued pre-training required—precisely the least reversible channel—to keep pace [13]. Knowledge conflicts are the second maintenance failure: Section 3.2's evidence that models defer inconsistently between fresh and parametric knowledge [10] means staleness is not merely missing information but actively misleading information the model may prefer. Hallucination is the downstream

symptom common to both: Ji et al.'s survey organizes the phenomenon—generation unfaithful to sources, whether contradicting them or fabricated from nothing—along with causes and mitigations, and its taxonomy has become the field's shared language for the defect [3].

TemporalWiki's measurement protocol is the part with the sharpest governance implication, because it shows the decay is not merely real but invisible to the affected party. The benchmark trains two models on adjacent Wikipedia snapshots and evaluates each on the other's snapshot, in both temporal directions, so that the measured quantity is precisely the knowledge difference between two moments of a changing world [13]. The finding that stings: a model degrades measurably on facts that changed after its training cutoff, yet nothing in its behavior flags the degradation—the model answers fluently and confidently from frozen premises, and only an external benchmark keyed to the world's own versioning reveals that the answers have gone stale [13]. Knowledge in parameters, in other words, decays on a schedule the operator cannot observe without constructing a diff against the world's current state; knowledge in an index decays exactly as fast as its sources do, and the diff is the index update itself. That asymmetry—identical decay, opposite observability—is arguably the strongest architectural argument for the retrieval channel in the entire governance literature, and it is invisible in benchmarks that evaluate capability at training time rather than knowledge maintenance over time [13].

The three maintenance failures are usually catalogued side by side, but the supply chain reading gives them a causal order that cataloguing obscures. Staleness originates upstream and midstream: the world changed and the injected knowledge did not, because the injection channel froze its contents at crawl or training time. Conflict arises next, in the gap between the frozen parameters and anything fresher the system consults—the model must adjudicate between its own aging knowledge and newer evidence, with the adjudication behavior documented as inconsistent. Hallucination compounds both: a model asked about a world its parameters no longer describe must either abstain—performance that benchmarks show is unreliable—or interpolate, and interpolation over stale premises is precisely where unfaithful generation concentrates. The ordering matters for governance because it locates the intervention points: staleness is addressed at injection architecture, conflict at adjudication behavior, and hallucination only at output—the last and weakest place to intervene. The maintenance record supports one further governance conclusion: knowledge inside parameters degrades on its own schedule, unobservable without dedicated benchmarks, while knowledge in retrieval degrades exactly as fast as its sources—and visibly. Whatever else the channels differ in, they differ in whether decay is auditable.

5. The Recall Problem: Removal Without a Mechanism

Recall is where the supply chain currently fails hardest, because the demand for it is legal while the capability is technical. TOFU, the first unlearning benchmark with knowable ground truth—fictitious author profiles that the model was trained on and can be probed against—reports that existing unlearning techniques preserve model utility while largely failing to remove the targeted knowledge, and that the failure worsens as the volume to forget grows [14]. The privacy stakes inherit from the memorization evidence of Section 3.3: content that cannot be verifiably removed remains extractable [11,12], whatever the data controller's policy says.

TOFU's design is what makes its negative result evidentially strong rather than merely discouraging. By training models on 200 fictitious authors—roughly 4,000 question–answer pairs generated for that purpose—the benchmark manufactures the condition real deployments lack: the ground truth of what the model knows and therefore of what "forgotten" would mean [14]. Its evaluation plane is correspondingly differentiated: beyond target knowledge, it probes forget quality (including membership-inference probes that ask whether the forgotten data remains detectable in the model's behavior), model utility on general and related author performance, and the trade-off frontier between the two. The reported pattern is the unlearning dilemma in miniature: methods that preserve utility largely fail to remove (the unlearned knowledge remains recoverable or inferable), and methods that remove more thoroughly pay for it in collateral degradation of unrelated capability—with the dilemma worsening as forget volume grows [14]. For governance, TOFU supplies two things at once: the demonstration that current unlearning is not deployment-grade, and the evaluation protocol that any future claim of verified deletion would have to survive—probing, inference attacks, utility accounting [14]. A deletion standard, in other words, already has a candidate form; what the field lacks is models that pass it.

The statutory side has more structure than the literature's references to "the right to be forgotten" suggest, and the structure matters for engineering. The GDPR's erasure right (Article 17) operates alongside the storage-limitation principle (Article 5), which requires that personal data not be kept longer than necessary—two distinct obligations, of which the second binds continuously, not only upon request—and the enforcement regime reaches global turnover percentages, which is what gives the obligations their pricing power [15]. The contested question for large models is threshold: whether personal data embedded in weights constitutes "processing" of that data at all, an interpretation on which the erasure obligation's applicability to trained models turns, and on which regulators and scholars have not fully converged [15].

But the uncertainty cuts against the field rather than for it: a controller who cannot verify what personal content sits in the weights—the memorization problem of Section 3.3—cannot establish that none is there, and the practical compliance posture therefore defaults to treating parameters as a plausible processing surface with no working deletion mechanism. That is the recall gap in statutory terms: two binding obligations, one contested threshold, and zero verified mechanisms on the technical side [14,15].

Law does not wait for the engineering. The GDPR's right to erasure obliges controllers to delete personal data, with narrow exemptions, and a corpus is personal data in the statute's terms long before it is weights [15]; the NYT complaint's requests go further, seeking to force unwinding of what training did [1]. The mismatch is stark: the legal system already issues recall orders against a supply chain whose recall stage has no verified mechanism. Corpus-side withdrawal—removing a source before training—remains the only recall that works as intended, and it acts only on models not yet built.

For models already deployed, the honest summary of the field's evidence is that recall is currently performed in three ways that work and one that is demanded but not yet delivered. The ways that work: drop the retrieval index, which removes the knowledge from the system's evidentiary base instantly; retrain from a cleaned corpus, which works but costs as much as building the model; and do not train on the content in the first place, the preventive form that opt-out signals, licensing negotiations, and provenance audits all serve. The way that is demanded but not delivered is surgical unlearning—removing one source's influence from finished weights—which is exactly the capability TOFU measures and finds wanting [14]. The pattern across the four is instructive: recall capability is inversely proportional to how deep the knowledge sits, which is Table 2's reversibility grade read backward. Governance that takes the evidence seriously will therefore concentrate on the two ends of that gradient—making the preventive end cheap enough to be used, and demanding verification at the surgical end rather than accepting its promises.

6. Governance Across the Chain: Control Points

The governance instruments that exist map naturally onto chain stages, which is both their strength and their limit. Upstream, datasheets for datasets propose the documentary record the chain lacks: motivation, composition, collection process, and recommended uses attached to every dataset, in the manner of component datasheets in electronics [16]—the instrument that, had it existed at scale, would have answered the audit questions that Dodge [5] and the Data Provenance Initiative [4] had to reconstruct after the fact. Downstream, model cards give

released models a documented boundary of intended use and evaluated performance [17]. The EU AI Act overlays both ends with enforceable obligations: training-data summaries and record-keeping for general-purpose model providers, and logging duties downstream for high-risk deployers [18].

The documentation instruments are also worth reading at the granularity of their questions, because their questions are the supply chain's questions. Datasheets specify a dataset's motivation, composition, collection process, preprocessing, and recommended uses—each heading a field the C4 audit found missing in practice: the collection process question is the provenance record the corpora lack; the preprocessing question is the record of what filtering removed; the composition question is what would have exposed the register skew before training rather than after [16]. Model cards, symmetrically, attach to a released model its intended use contexts, evaluation data, and performance broken down by demographic and task group—a reporting structure whose disaggregation discipline is what makes decline-on-subpopulations visible instead of averaged away [17]. The two instruments bracket the chain at its documentable ends, which is exactly their limit: between the dataset that enters and the model that ships sit the injection channels, and no instrument in current standard use requires stating which channel carries which knowledge or how reversible that carriage is. The documentation tradition, in other words, has already solved the hardest problem—making disclosure habitual—and its unfinished work is extending the disclosure across the midstream [16,17].

Two gaps remain that documentation cannot close. First, the midstream lacks its control point entirely: no standard requires a deployment to state which channel carries which knowledge class, though Table 2's reversibility grading is exactly the kind of statement that could be required—a system carrying personal or licensed knowledge in the least reversible channel should have to say so. The AI Act's training-data summary obligation for general-purpose providers is the nearest existing instrument, and its formulation—a public "sufficiently detailed summary" of training content—will be tested against precisely this question: whether a summary that omits provenance quality and channel reversibility can be considered sufficiently detailed at all [18]. Second, recall lacks verification: deletion claims are currently unfalsifiable without probing protocols of the kind TOFU introduced [14], and a governance regime that accepts unaudited deletion claims replicates upstream's original sin—accepting inputs on faith. The supply chain reading turns these from philosophical worries into specification gaps with natural owners: standards bodies for the documentation tiers, model providers for reversibility

disclosure, deployers for verified deletion.

The AI Act's obligations, read closely, already sketch the midstream disclosure this survey argues for, though in entry-event rather than channel terms. The Act's documentation duties for general-purpose model providers include technical documentation covering the training stage and a publicly posted "sufficiently detailed summary" of training content—the first time European law has required model developers to describe what went into the weights at all [18]. Its record-keeping obligations (automatically generated logs for high-risk systems) and post-market monitoring duties extend the governance horizon from authorization to lifecycle, which is the same reorientation the recall problem demands: a model whose knowledge can be corrected or removed is governed by a different regime than one whose knowledge is frozen at release. What the Act does not yet do—the gap this survey's Table 2 identifies—is require the channel-level statement: which knowledge classes entered through which injection channel, and what each channel's reversibility grade implies for the operator's correction and deletion obligations. That omission is where the specification work remains: the Act created the disclosure habit; the supply chain framework specifies what the disclosure must say to govern knowledge rather than merely document it [18].

7. Conclusion

This survey organized the knowledge governance of large language models as a supply chain and reached three findings. First, the upstream is structurally unauditable today: the field's own audits document corpora with no provenance discipline [5] and datasets that mostly fail to convey their licensing terms [4]. Second, the midstream's four injection channels are graded—retrieval fully reversible [6], editing locally reliable but globally leaky [7-9], pre-training effectively irreversible [11]—and deployment practice inverts the governance order, concentrating knowledge where it is hardest to govern. Third, recall is demanded by law but not delivered by technique: unlearning remains fragile at scale [14], erasure obligations bind regardless [15], and litigation has arrived to test the gap [1]. The control points that close these gaps—provenance records [16], reversibility disclosure, verified deletion—are documentary and procedural rather than heroic, which is precisely why they are achievable.

The sequencing of that work is not arbitrary, because the control points depend on each other in one direction. Verified deletion is the glamorous problem, but it is downstream of reversibility disclosure—an operator cannot be held to delete what no record says was injected—and reversibility disclosure is downstream of provenance documentation, since "which knowledge entered" presupposes "what was known about what entered." The

dependency chain mirrors the stages themselves, which is the deepest sense in which the supply chain is the right frame: upstream defects do not stay upstream, they propagate into every later control's preconditions, exactly as a defective input lot propagates through a manufacturing process. Governance programs that begin at recall—demanding deletion first—therefore invest at the point of maximum dependency and minimum feasibility, while programs that begin at documentation make every later stage cheaper and more verifiable. The recommendation this survey's evidence supports is accordingly ordered: make the documentary habit universal at the upstream [16,17], extend it across the midstream as channel-level reversibility disclosure [18], and let the demand for verified deletion [14,15] concentrate the field's research investment where the engineering is genuinely missing.

The framework travels to any technology whose output is learned rather than recorded—recommender systems, foundation-model derivatives, embodied controllers. For language models specifically, the highest-value open problems follow the chain: upstream, licensing infrastructure that makes permission machine-checkable; midstream, ripple-propagating edit methods that would move model editing from Table 2's "partial" grade toward the retrievable column; downstream, maintenance benchmarks that treat staleness as a reportable property; and at recall, deletion that survives adversarial probing. The NYT lawsuit asked a court to do what the technology cannot; building the technology that can is the field's share of the answer.

References

[1] The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. 2023).

[2] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[3] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, Y., Xu, D., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[4] Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2023). The Data Provenance Initiative: A large scale audit of dataset licensing & attribution in AI. arXiv preprint arXiv:2310.16787.

[5] Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information

Processing Systems 33.
[7] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems 35.
[8] Meng, K., Sen Sharma, A., Andonian, A., Belinkov, Y., & Bau, D. (2023). Mass-editing memory in a transformer. In The Eleventh International Conference on Learning Representations (ICLR).
[9] Zhong, Z., Wu, Z., Manning, C. D., Potts, C., & Chen, D. (2023). MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
[10] Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In The Twelfth International Conference on Learning Representations (ICLR).
[11] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633-2650).
[12] Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022). What does it mean for a language model to preserve privacy? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 2280-2292).
[13] Jang, J., Ye, S., Lee, C., Lee, S., Yoon, D., Seo, M., & Yang, S. (2022). TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.
[14] Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). TOFU: A task of fictitious unlearning for LLMs. In First Conference on Language Modeling (COLM).
[15] European Parliament and Council. (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). Official Journal of the European Union, L 119.
[16] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
[17] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*) (pp. 220-229).
[18] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.