Analysis

When Training Needs Permission and Retrieval Is Metered: The Open Web Is Turning from Free Feedstock into a Metered Input

HX Research

HX Research · LLM Technical Brief · November 4, 2026

For years, large language models have rested on an unstated assumption: the open web can be crawled, trained on and searched at near-zero marginal cost. That assumption is now loosening at both ends. On the training side, major model developers have said they will honour a shared robots.txt training-preference signal based on the IETF AI Preferences working group's drafts, and measurement studies show that restrictions concentrate on high-quality, actively maintained sources. On the inference side, search and fetch tools are now billed per successful call, and the retrieved content is billed again as input tokens. Existing discussions treat these shifts separately and usually measure losses by token share, which understates the problem. We argue that the open web is becoming a metered input: the marginal source of model capability is shifting from crawl scale to rights-aware data mixtures and budgeted retrieval, and enterprises should manage data rights and retrieval calls as first-class cost items. We formalize quality-weighted restriction, data-constrained scaling, per-call retrieval cost, budgeted gating and the internalize-versus-retrieve break-even, and contribute a framework, a route comparison, an algorithm, seeded illustrative calculations and five falsifiable forecasts.

Keywords: AI training preferences; robots.txt; data licensing; metered retrieval; data-constrained scaling; synthetic rewriting; budgeted retrieval

Introduction

For the past five years, the data foundation of LLM pretraining has come almost entirely from one source: large-scale crawling of the open web. Public archives such as Common Crawl[1], after cleaning, deduplication and quality filtering, became successive generations of corpora — C4, The Pile, RefinedWeb, Dolma, FineWeb and DCLM[2–8]. Scaling laws supplied the economic rationale: under a fixed compute budget, parameters and tokens should grow together, and the default source of extra tokens was simply to crawl more[9,10]. In that narrative, web pages were essentially free feedstock and the real cost lay in compute.

That default is now loosening on the training side. A longitudinal audit of 14,000 web domains by Longpre and colleagues found that within a single year (2023–2024), robots.txt restrictions fully restricted about 5%+ of all tokens in C4 and 28%+ of its most actively maintained, critical sources; counting terms-of-service restrictions, 45% of C4 was restricted (per the paper)[11]. The same group's earlier licensing audit of more than 1,800 text datasets reported license omission rates above 70% and error rates above 50% on widely used dataset-hosting sites (per the paper)[12]. The distribution of restrictions is even more telling. A study of news websites found that 60.0% of reputable sites disallow at least one AI crawler, compared with 9.1% of misinformation sites, and that blocking by reputable sites rose from 23% in September 2023 to nearly 60% by May 2025 (per the paper)[13]. Another study covering the top one million websites reported that 55% of news outlets with high factual reporting disallow GPTBot, against only 4.1% of right-leaning outlets (per the paper)[14]. Surveys and measurements of creators show that the efficacy of robots.txt is limited by technical awareness and by whether crawlers comply; blocking at the reverse-proxy layer is stronger but sparsely deployed[15].

On 30 October 2026 the trend entered a standardization phase. OpenAI announced that it will honour a common robots.txt signal for foundational generative-AI training, aiming to implement it in the coming months while its existing crawler controls continue to apply[16]. Google said that when a site uses the shared 'policy token' to opt out of training without explicitly naming Google-Extended, it will treat this as a disallow of Google-Extended, and that it aims to support the control in 2027[17]. According to press reports, Anthropic has already adopted the signal, Microsoft plans to support it next year, and Meta signalled support without committing to honour it[18]. The semantics of these signals come from the vocabulary and attachment drafts of the IETF AI Preferences (AIPREF) working group[19,20], whose charter explicitly places technical enforcement of preferences, client authentication and auditing of AI training out of scope[21]. Training preferences, in other words, are becoming expressible and uniform, but their binding force still rests mainly on voluntary compliance and regulatory

pressure.

A parallel shift is under way at inference time, and it shows up more directly as a price. DeepSeek's official documentation states that its Search and Fetch APIs are billed per successful request at US$0.75 per 1,000 calls each, and that server-side search and fetch calls triggered when its API is used through DeepSeek Harness, Claude Code or Codex are billed from 00:00 Beijing time on 1 November 2026[22]. Anthropic's web search tool costs US$10 per 1,000 searches, with retrieved content billed separately as input tokens and carried into later turns[23]; OpenAI's price list likewise separates web-search tool calls from search-content tokens, charging some models a fixed block of 8,000 input tokens per call[24]. On the publisher side, Cloudflare has piloted pay-per-crawl since 2025, negotiating with crawlers through HTTP 402 responses and a price header[25]; researchers have proposed protocols such as terms.txt, which declares per-path, per-purpose machine-access terms backed by signed receipts[26], as well as pricing agents that set pay-per-crawl prices for heterogeneous content[27].

Why does retrieval pricing matter now? Because agents are turning retrieval from an occasional lookup into a continuous background load. The Super-agent design described by the DeepSeek Harness team is a persistent, event-driven agent that spans sessions and advances several tasks in parallel; the team says internal instances already run continuously on messaging and code-hosting platforms[28]. A characterization of production agent workloads found heavy cross-request redundancy in search queries and web fetches across three applications, with tool-result caching removing 35.2% of redundant search calls (per the paper)[29]. Once the number of retrieval calls scales with an agent's uptime rather than with the number of user questions, per-call billing stops being rounding error and becomes a structural cost.

Existing discussions leave three gaps. First, training-side restrictions are usually reported as token shares, but restriction correlates positively with source quality, so token-share figures systematically understate the effect on model quality[11,13,14]. Second, research on data-constrained scaling, on scaling with synthetic rewrites and on scaling retrieval-augmented models belongs to separate communities and is rarely placed in a single cost ledger[30–33]. Third, enterprises almost never treat "should this knowledge be trained into the model or looked up each time?" as a computable decision — yet once retrieval is metered, that decision has a well-defined break-even point.

Our central thesis is that the open web is turning from free feedstock into a metered input: training is constrained by declarable preference signals and inference by per-call billing; because restrictions concentrate on high-quality sources, the marginal source of model capability is shifting from crawl scale to rights-aware data mixtures and budgeted retrieval, and enterprises should manage data rights and retrieval calls as first-class cost items on a par with compute. We make three contributions:

1. A unified formalization (equations (1)–(6)) of quality-weighted restriction, effective data under constraint, per-call retrieval cost, budgeted gating and the internalize-versus-retrieve break-even, with a notation table. It shows why token-share restriction understates losses and how retrieval prices decide whether knowledge belongs in parameters or in retrieval.
2. A framework of four input channels × three control signals, and a mechanism-level comparison of five routes — crawling at scale, licensed and openly licensed corpora, synthetic rewriting, inference-time retrieval, and private data with small domain models — with evidence for and against each.
3. A rights-aware data and retrieval budgeting algorithm (Algorithm 1), seeded illustrative calculations, and five forecasts, each with a time horizon, evidence, falsification condition and enterprise implication.

The paper is organized as follows. Section 2 gives the formalization and notation; Section 3 the framework; Section 4 the mechanism analysis and route comparison; Section 5 the algorithm; Section 6 the illustrative calculations; Sections 7–10 cover forecasts and a research agenda, common misconceptions, limitations and conclusions.

Problem formalization and notation

Quality-weighted restriction

Let the candidate training corpus consist of domains (or sources) d = 1,...,D. Source d contributes n_d tokens, has quality weight $w_d \in [0,1]$ (its marginal contribution to downstream capability, often approximated by a model-based quality classifier[7,8]), and restriction indicator $o_d \in \{0,1\}$ (training is disallowed by robots.txt, an AI-preference signal, a TDM reservation or the terms of service). The token-weighted and quality-weighted restriction rates are

$$\rho_{tok} = \sum_d n_d o_d / \sum_d n_d, \qquad \rho_q = \sum_d n_d w_d o_d / \sum_d n_d w_d \tag{1}$$

Taking expectations with token weights gives a simple but important identity:

$$\rho_q - \rho_{tok} = Cov_n(w, o) / E_n[w] \tag{2}$$

Equation (2) says that whenever restriction correlates positively with quality ($Cov_n(w,o) > 0$), the token-share figure understates the cut to effective data, by exactly the covariance divided by mean quality. The audits cited above — 5% of all tokens versus 28% of critical sources, 55% of high-factual news outlets versus 4.1% of right-leaning ones — are empirical signs that this covariance is positive (per the papers)[11,14]. The indicator o_d can be generalized to a compliance probability: preference signals carry no built-in enforcement[21,34], and providers' implementation timelines differ[18], so the data that actually enter training pass through two filters — declaration and compliance.

Effective data under constraint

Restriction directly reduces the number of unique tokens U. Muennighoff and colleagues model the declining value of repeated data as an exponential decay; after R repetitions the effective data are

$$D' = U + U \cdot R^* \cdot (1 - exp(-R/R^*)), \qquad L(N, D') = E + A/N^{\alpha} + B/D'^{\beta} \tag{3}$$

where R* is the decay constant for the value of repetition and L takes the Chinchilla form[10,31]. According to the paper, with a fixed compute budget, up to about four epochs of repetition yield negligible changes in loss relative to unique data; with more repetition the value of added compute eventually decays to zero[31]. Equation (3) implies that restriction first compresses U, that repetition can compensate only up to a ceiling, and that — because U shrinks more in its high-quality part (equation (2)) — what really needs replenishing is $\sum n_d w_d$, not $\sum n_d$.

Per-call retrieval cost

The per-task cost of inference-time retrieval splits into generated tokens, retrieval calls and back-filled retrieved content. Let tool k (search, fetch and so on) cost p_k per successful call, be called m_k times, and back-fill $T_{r,k}$ input tokens per call; with input and output token prices c_{in} and c_{out},

$$c_{task} = c_{in} T_{in} + c_{out} T_{out} + \sum_k m_k (p_k + c_{in} T_{r,k}) \tag{4}$$

The bracketed term $p_k + c_{in} T_{r,k}$ is the effective price per retrieval, p_{eff}. The decomposition has an often-overlooked corollary: because retrieved content is billed as input tokens and may occupy the context again in later turns[23,24], back-filled tokens can dominate retrieval cost even when the call itself is cheap; and since the position of evidence in a long context affects how well it is used, stuffing in more is not the same as using more[35].

Budgeted gating

For a query x, let a(x) be the probability of a correct answer without retrieval and r(x) the probability with retrieval; the marginal utility of retrieval $\Delta(x) = r(x) - a(x)$ can be negative (retrieval harm). A gate $g(x) = 1[u(x) \geq \tau]$ decides whether to retrieve on the basis of a computable uncertainty score u(x), with the threshold τ set by a budget b:

$$max_{\tau}\ E[a(x) + g(x)\Delta(x)] \quad s.t. \quad E[g(x)] \leq b, \qquad C_{ret} = E[g(x)] \cdot \bar{m} \cdot p_{eff} \tag{5}$$

This formulation matches budget-aware evaluation that recasts active retrieval as utility estimation: retrieval is valuable only through its marginal correctness change over a no-retrieval answer, and nominal thresholds may miss target usage on future data (per the paper)[36]. Training-free gating from the logits of a short draft[37], reinforcement learning that separates the decision to search from problem solving[38], search agents that recognize their own knowledge boundary[39] and tool-efficiency rewards[40] can all be read as different implementations of u(x) or g(x); Self-RAG and Adaptive-RAG, which routes by question complexity, were early forms of the idea[41,42].

The internalize-versus-retrieve break-even

A body of knowledge can be trained into the parameters (licensed, curated, trained or distilled, and refreshed as the knowledge changes) or looked up each time. Let the one-off internalization cost be $C_{lic} + C_{train}$, the cost per refresh C_{ref}, the knowledge half-life τ_k, the planning horizon H and the mean number of retrievals per query $\bar{m}$. Internalizing beats retrieving when

$$Q_H \cdot \bar{m} \cdot p_{eff} > C_{lic} + C_{train} + C_{ref} \cdot H/\tau_k, \qquad i.e. \qquad Q^* = (C_{lic} + C_{train} + C_{ref} H/\tau_k) / (\bar{m} \cdot p_{eff}) \tag{6}$$

Equation (6) turns the familiar enterprise question ”should this knowledge go into the model?” into a computation: the larger the query volume Q_H, the more expensive retrieval and the more stable the knowledge (larger τ_k), the stronger the case for internalizing; volatile knowledge and sparse queries favour retrieval. This is consistent with the retrieval-augmented literature — RETRO, Atlas and MassiveDS show that larger datastores can trade for fewer parameters at equal or better performance on knowledge tasks[33,43,44] — equation (6) simply places those findings inside a budget constraint that includes prices and licences.

Notation

Table 1 summarizes the notation.

Analytical framework: four input channels and three control signals

Placing equations (1)–(6) in a system view, the channels through which a large model acquires knowledge from the outside world fall into four groups: crawling the open web at training time, licensed and openly licensed corpora at training time, synthetic and rewritten data at training time, and retrieval at inference time. Each channel is acted on by three kinds of control signal: declaration (what the content owner allows or disallows), pricing (how much each unit of access costs) and enforcement (who ensures that declarations and prices are honoured). Table 2 sets out the framework.

The three signals must be separated because they are maturing at different speeds. Declaration is standardizing quickly: the IETF vocabulary distinguishes usage categories such as AI Training (train-ai, using an asset to modify a model’s learned parameters), AI Use (ai-use, using an asset as input to a generative AI model) and Search (search, applications whose primary purpose is to select assets and direct users to their location), and states that a recipient processes only the expression methods it has implemented[19]; the attachment draft defines how these preferences are carried in HTTP and robots.txt[20]. TDMRep, which predates AIPREF, addresses rights reservation under the EU text-and-data-mining exception[45]. Pricing is still at the pilot and proprietary stage: publisher-side pay-per-crawl, tool-side per-call search, and academic proposals for HTTP 402 negotiation with signed receipts[22,25,26]. Enforcement is the weakest layer: the robots.txt specification itself states that ”these rules are not a form of access authorization”[34], and the AIPREF charter excludes techni-

Table 1 | Main notation

Symbol	Meaning	Where
n_d, w_d, o_d	Source tokens, quality weight, restriction indicator (or compliance probability)	Eqs (1)–(2)
ρ_{tok}, ρ_q	Token-weighted and quality-weighted restriction rates	Eqs (1)–(2)
U, R, R*, D′	Unique tokens, repetitions, repetition decay constant, effective data	Eq. (3)
p_k, m_k, $T_{r,k}$, p_{eff}	Tool price, number of calls, back-filled tokens, effective price per retrieval	Eq. (4)
a(x), r(x), Δ(x), u(x), τ, b	Correctness without/with retrieval, marginal utility, uncertainty score, threshold, budget	Eq. (5)
C_{lic}, C_{train}, C_{ref}, τ_k, Q*	Licence, training and refresh costs, knowledge half-life, break-even query volume	Eq. (6)

Table 2 | Four input channels × three control signals: representative mechanisms, main risks and observable indicators

Channel	Declaration	Pricing	Enforcement	Main risks	Observable indicators
Open-web crawling (training)	robots.txt, AIPREF train-ai, TDMRep, terms of service	Mostly zero; pay-per-crawl pilots	Voluntary compliance plus law (EU TDM opt-out)	Quality-weighted loss exceeds token loss; source skew	ρ_{tok}, ρ_q, restriction by quality quantile
Licensed and openly licensed corpora (training)	Licence text and data cards	One-off licence fee or zero	Contracts and licences	Faulty licence metadata; limited scale and diversity	Licence-verification coverage, corpus diversity
Synthetic and rewritten data (training)	Generator licence and terms of use	Generation compute	Internal governance	Model collapse, narrowed distributions, inherited generator bias	Rewrite share, retained real-data share, verifier pass rate
Inference-time retrieval	AIPREF search and ai-use, site terms	Per-call price plus back-filled tokens	Platform billing and site blocking	Cost grows with agent uptime; retrieval harm	Calls per task, realized retrieval rate, p_{eff}

cal enforcement[21]. Real constraint comes mainly from law and compliance pressure. Article 53 of the EU AI Act requires providers of general-purpose models to put in place a policy to comply with Union copyright law, to identify and comply with rights reservations under the TDM exception, and to publish a summary of training content[46], and the accompanying General-Purpose AI Code of Practice contains a copyright chapter[47]. China's Interim Measures for the Management of Generative AI Services require, in Article 7, the use of data and foundation models from lawful sources and prohibit infringing others' intellectual-property rights[48].

The mismatch between layers has two direct consequences. First, the same page is usable by different providers at different times, so the snapshot date of a training corpus becomes part of the recipe. Second, finer-grained declarations allow training and retrieval to be treated separately: a site can permit search while forbidding training (for example, train-ai=n, search=y), which pushes its knowledge out of the parameters and into retrieval — and hence into the priced ledger of equations (4)–(6). The next section examines the mechanisms channel by channel.

Mechanism analysis and route comparison

Training side: semantics, force and quality skew of preference signals

The first mechanical effect of preference signals is to replace site-by-site naming of crawlers with a single declaration that covers every participating party. OpenAI framed the benefit as sparing website operators the work of maintaining separate rules for each provider[16]; Google gave a concrete interpretation rule: a shared token that opts out of training, with Google-Extended not explicitly specified, is treated as a disallow of Google-Extended[17]. Once the cost of declaring falls, the number of declarations should rise — and it should rise fastest among sources with editorial teams, legal resources and high-quality content.

The second effect is quality skew. The covariance term in equation (2) is not a theoretical construct. The gap in AI-crawler blocking between reputable and misinformation news sites, the positive correlation between factual reporting and blocking, and the finding that politically neutral outlets restrict most (58%) while hyperpartisan and low-factual sites restrict less (per the papers)[13,14] all point the same way: a crawler that honours preference signals sees a web tilted towards lower-quality and more polarized content. Longpre and colleagues warn accordingly that, if respected or enforced, these restrictions will rapidly bias the diversity, freshness and scaling laws of general-purpose AI systems (per the paper)[11]. There is a temporal skew too: analyses of open pretraining corpora find that a model's effective knowledge cutoff often differs from its reported cutoff, partly because new Common Crawl dumps contain substantial amounts of old data (per the paper)[49]. When

fresh high-quality content is blocked while old content stays in the snapshots, freshness declines further.

The third effect concerns memorization and copyright risk, which in turn strengthens content owners' motive to block. Early work showed that verbatim training data can be extracted from large models[50], and that memorization grows log-linearly with model capacity, the number of times an example is duplicated and the length of the prompting context (per the paper)[51]; measurements on 14 open-weight models and 200 books show that most models do not memorize most books, but that there are notable exceptions (per the paper)[52]. This structure of evidence means the rights question cannot be dismissed with "the risk is small on average", and it makes content owners more inclined to declare first and negotiate later.

Filling the gap with repetition, rewriting and licensed corpora

There are three ways to fill the training-side gap, each with its own mechanism and limits.

The first is repetition. Equation (3) shows that modest repetition costs almost nothing but yields diminishing returns[31]. Recent work refines this conclusion. Baron and colleagues note that past data-constrained prescriptions held hyperparameters such as weight decay fixed with scale, which can make loss rise with compute; transferring hyperparameters with scale and adding ensembling reduced validation loss by more than 0.15 nats on 100M FineWeb tokens (per the paper)[53]. Another study finds that the ranking of repetition counts changes with model size; in a 520M-parameter Proof-Pile-2 experiment, cutting repetition from sixteen to eight epochs improved loss (per the paper)[54]. When data are scarce and compute plentiful, masked diffusion language models make better use of repeated data than autoregressive models (per the paper)[55]. The common implication is that, under restriction, how many passes to make over the same data — and how — becomes a recipe variable that must be calibrated to scale.

The second is synthesis and rewriting. Phi-1 combined about 6B tokens of "textbook-quality" web data with about 1B tokens of GPT-3.5-generated textbooks and exercises to train a 1.3B-parameter code model reaching 50.6% pass@1 on HumanEval (per the paper)[56], showing that high-quality synthetic data can stand in for large volumes of raw data. Roberts and colleagues' rewriting scaling law, which charges the compute spent generating rewrites, finds that the optimal rewrite fraction rises with compute, from under 10% at 3×10^{18} FLOPs to 32% at 10^{22} FLOPs; in a data-limited setting with only a fixed seed corpus, a model trained on rewrites of its own pretraining data, produced by a rewriter of the same size, beats the best natural-data-only run at every size up to 3B parameters (per the paper)[32]. Synthetic data have known failure modes, however. Shumailov and colleagues reported in *Nature* that indiscriminately training on model-generated content causes irreversible defects in which the tails of the distribution disappear[57]; self-consuming loops without enough fresh real data lose quality or diversity generation by generation[58]. The counter-evidence is equally clear: if synthetic data accumulate alongside real data rather than replacing them, test error has an upper bound independent of the number of iterations[59]; an external verifier can prevent collapse, but in the long run drives the model towards the verifier's own "knowledge centre", and with an imperfect verifier early gains plateau or reverse[60]. In terms of equations (1)–(3), synthetic rewriting amplifies the high-quality U that remains; it does not recover information unique to blocked sources.

The third is licensed and openly licensed corpora. A best-practices report on open datasets published in January 2025 still noted that no model had yet been trained at meaningful scale on open-access and public-domain data alone[61]; in June the same year, Common Pile v0.1 assembled about 8 TB of public-domain and openly licensed text from 30 sources, and two 7B-parameter models trained on it performed competitively with Llama 1 and Llama 2 7B at similar compute (per the paper)[62]. Code had precedents earlier: The Stack includes only permissively licensed source code and provides an "Am I in The Stack" lookup and removal process[63]; its second generation is built on the Software Heritage archive[64]. These efforts show that licence-clean corpora can reach usable scale. Their limits are diversity and freshness — news, industry reports and professional-community content are precisely what is least often released under open licences.

Inference side: retrieval moves from free extra to metered service

The original motivation for retrieval-augmented generation was to separate knowledge from parameters so that it could be updated and attributed[65,66]. RETRO, with a 2-trillion-token database, matched GPT-3 on The Pile with about 25× fewer parameters; Atlas exceeded 42% accuracy on Natural Questions with only 64 examples, beating a 540B-parameter model by 3 points with 50× fewer parameters; and MassiveDS, a 1.4-trillion-token datastore, showed that a smaller model with a large datastore can outperform a larger LM-only model on knowledge-intensive tasks (all per the papers)[33,43,44]. WebGPT, Toolformer, ReAct and Search-R1 then turned retrieval from a fixed pipeline into an action the model chooses[67–70], and benchmarks such as BrowseComp measure persistent search for hard-to-find information as a capability in its own right[71].

The route carried an implicit premise: retrieval calls are cheap, and those being searched are willing to be searched. Both premises are changing at once. On price, p_k in equation (4) is now an explicit line in public price lists, and it can differ by more than an order of magnitude between providers[22,23]. On availability, a study of 11,500 real queries found that sites blocking Google's AI crawler are significantly less likely to be cited by AI Overviews, and that generative and traditional search retrieve very different sources (average Jaccard similarity below 0.2) (per the paper)[72]; work on the digital commons warns that if people turn to generative AI for information and stop contributing, public knowledge may become undersupplied

and the open web may close through access restrictions[73]. Retrieval, in short, cannot unconditionally substitute for blocked training data: the AIPREF vocabulary lets content owners declare training, AI use and search separately[19], and pay-per-crawl lets "searchable" be priced as well[25,27].

Price divergence also changes what evaluations mean. BrowseComp relies on live web-search APIs, whose dynamic and opaque behaviour hampers fair comparison between deep-research methods; BrowseComp-Plus switches to a fixed corpus with human-verified supporting documents in order to isolate the retriever's contribution[74]. Once retrieval is metered, two systems with equal accuracy but a threefold difference in calls are, for an enterprise, two entirely different products.

When to retrieve: gating, budgets and retrieval harm

The crux of equation (5) is that $\Delta(x)$ can be negative. Budget-aware evaluation of active retrieval finds that retrieval harm is non-negligible, that router rankings change across datasets and budgets, that nominal thresholds can miss target usage, and that simple uncertainty or retrieval-score baselines often rival learned utility routers (per the paper)[36]. That gives enterprises a pragmatic starting point: first bring the retrieval rate down with a cheap gate, then consider sophisticated learned routing.

Current methods implement gating at three levels. At inference, TARG computes uncertainty from the prefix logits of a short, context-free draft and retrieves only above a threshold, with no extra training[37]. In training, AdaSearch uses a two-stage, outcome-driven reinforcement-learning framework that decouples problem solving from the decision to search, substantially reducing unnecessary search calls at a small accuracy cost relative to always searching (per the paper)[38]; IKEA uses a knowledge-boundary-aware reward so that the agent prefers internal knowledge[39]; and ZEST-Search reduces redundant searches with a tool-aware reward in a zero-label setting[40]. At the system level, caching and deduplication act directly on $\bar{m}$: the heavy redundancy in production search and fetch traces means that a sizeable share of retrieval spending need not be incurred[29].

Counter-evidence: why pessimism is unwarranted

Our thesis is easily misread as "the open web is running dry", so counter-evidence deserves equal space. First, forecasts of the data stock carry wide intervals: Villalobos and colleagues estimate that, if trends continue, training datasets will approach the stock of public human text between 2026 and 2032 (per the paper)[30], while noting that synthetic data, transfer from data-rich domains and data-efficiency gains could sustain progress. Second, advances in filtering and deduplication keep raising the value of each token: RefinedWeb showed that properly filtered and deduplicated web data alone can train strong models[5], and FineWeb and DCLM showed that model-based quality filtering is key to a high-quality training set[7,8]. Third, openly licensed corpora and rewriting scaling laws already offer workable paths[32,62]. Fourth, preference signals are honoured voluntarily and on scattered timelines[18,21], so their short-term effect on released models is limited.

This counter-evidence does not overturn the thesis; it calibrates its strength. The problem is not that data are disappearing, but that free high-quality data are shrinking in a way token shares cannot reveal — and that every way of filling the gap, whether repetition, rewriting, licensing or retrieval, carries a measurable price.

Table 3 compares the five routes side by side.

Method: rights-aware data and retrieval budgeting

We condense the analysis into an executable procedure. The goal of Algorithm 1 is to decide, under given compute, licensing and retrieval budgets, which knowledge enters the training mixture and which is left to retrieval, and where to set the retrieval gate — while leaving an auditable rights ledger.

Algorithm 1 embodies three design choices. First, step 1 treats preference state as time-varying data rather than a one-off compliance tick box, echoing the snapshot-time biases revealed by Dated Data[49]. Second, the order of step 3 reflects the counter-evidence: licensed corpora come first because they solve quality and auditability together; rewrites are appended, never substituted, because accumulation avoids collapse and replacement does not[57,59]. Third, step 5 requires reporting realized usage and harm rates rather than accuracy alone, as budget-aware retrieval evaluation recommends[36].

Illustrative calculations

All numbers below are illustrative; they show how equations (1)–(6) interact. Scripts and the random seed (20261104) are in src/figs.py; they do not represent measurements of any real system.

Quality-weighted restriction. We simulate 20,000 sources: token counts are log-normal, quality weights follow Beta(2,3), and the probability of restriction rises with quality through a logistic function. The token-weighted restriction rate ρ_{tok} is 11.3% and the quality-weighted ρ_q is 16.9%, a gap of 5.6 percentage points that matches the right-hand side of equation (2), $\mathrm{Cov}_n(w,o)/\mathrm{E}_n[w]$. By quality decile, restricted tokens make up about 0.5% of the lowest decile and about 43.0% of the highest (Fig. 1). The same report of "roughly a tenth of the data restricted" can thus correspond to losing about a sixth of the effective data and more than two-fifths of the top sources.

The internalize-versus-retrieve break-even. Assume a one-off licence-plus-training cost of US$20,000, US$2,000 per refresh, a 12-month horizon, three retrievals per query, 3,000 back-filled input tokens per retrieval and an input-token price of US$0.30 per million (illustrative). We take two public price points for retrieval calls: US$0.75 and US$10 per 1,000[22,23]. The effective price per retrieval p_{eff} is then US$0.00165 and US$0.0109 respectively; notably, at the lower price point the back-filled tokens (US$0.0009) already cost more than the call itself. By equation (6), with

Table 3 | Mechanism comparison of five routes to data and knowledge (evidence mostly from papers or official sources)

Route	Source of capability	Marginal cost structure	Quality and skew risk	Freshness	Compliance auditability	Best fit
R1 Crawling at scale with preference filtering	Scale of the open web	Crawl and filter compute, near-zero licence fees	Restriction correlates with quality; skew towards low-quality and polarized content	Snapshot-dependent; effective cutoff often lags	Requires retaining preference state at snapshot time	Base layer of general foundation models
R2 Licensed and openly licensed corpora	Licence-clean text and code	One-off licence or curation cost	Limited diversity and timeliness	Low to medium	High (licences and data cards)	Legally sensitive settings; models with public recipes
R3 Synthesis and rewriting	Amplifies retained high-quality data	Generation compute (computable via scaling laws)	Collapse and narrowing; inherited generator bias	Depends on seed data	Medium (log generator and share)	Data-limited, compute-rich
R4 Inference-time retrieval and search agents	External real-time information	Per-call price plus back-filled tokens	Retrieval harm; availability changes with declarations	High	Medium to high (citations and calls can be logged)	Volatile knowledge, low query volume
R5 Private data with small domain models and in-house retrieval	Enterprise documents and data	Internal curation and training; low marginal retrieval cost	Narrow coverage	Set by the enterprise's update cadence	High (data stay on premises)	Stable internal knowledge, high query volume

Algorithm 1 | Rights-Aware Data and Retrieval Budgeting (RADRB)

Input: Candidate sources {d} (tokens n_d, quality score w_d, robots.txt/AIPREF/TDMRep/terms status at crawl time); inventory of licensed and openly licensed corpora; knowledge needs {k} (expected query volume Q_k, half-life τ_k); tool prices $\{p_k\}$, back-fill token price c_{in}; compute, licensing and retrieval budgets b

Output: Training mixture (natural/licensed/rewritten shares), knowledge partition (internalize or retrieve), gate threshold τ with expected retrieval rate, rights and call ledger

1: Build the source ledger: for each d record crawl time, the verbatim preference declaration and its hash, and each provider's implementation status; record o_d as a compliance probability, not a Boolean
2: Compute ρ_{tok} and ρ_q by equations (1)–(2) and report restriction by quality quantile; flag "quality skew" if $\rho_q - \rho_{tok}$ exceeds a preset threshold
3: Fill the gap: first replenish high-quality U with licensed and openly licensed corpora; then reach D′ with scale-calibrated repetition per equation (3); finally, compute permitting, add rewritten data — appended, never substituted for real data — and spot-check with a verifier
4: Partition knowledge: for each k compute Q* by equation (6); internalize (train, distil or pre-index in the enterprise knowledge base) where $Q_k > Q^*$ and τ_k is long; leave the rest to retrieval
5: On a held-out set choose the gate threshold τ by equation (5) so that the realized retrieval rate stays within b; report retrieval-harm rate, threshold-transfer error and calls per task $\bar{m}$
6: Enable result caching and query deduplication on the retrieval path; recompute p_{eff} and $\bar{m}$ weekly
7: Re-run steps 1–6 monthly or whenever preference signals or price lists change; version the mixture, threshold and ledger

Complexity and applicability: Steps 1–2 are O(D); step 4 is O(K); step 5 sorts Q held-out queries in O(Q log Q). Suited to enterprise deployments with known query volumes and price lists; for sources whose quality w_d cannot be estimated, fall back to token weighting and say so in the report

a knowledge half-life of one month the break-even volume is about 8.89 million and 1.35 million queries a year; with a 12-month half-life it falls to about 4.44 million and 0.67 million (Fig. 2). The more expensive retrieval and the more stable the knowledge, the sooner internalizing pays.

Budgeted gating. We simulate 10,000 queries: correctness without retrieval follows Beta(2.2,1.6); retrieval harms 8% of queries and helps the rest by varying amounts; the gate uses a noisy uncertainty score. Never retrieving gives 58.0% accuracy; always retrieving gives 69.4% at about US$32.7 per 1,000 queries in retrieval fees (US$10 per 1,000 calls, three calls each). With a 30% retrieval budget, the uncertainty gate reaches 64.1% at about US$9.8, and an oracle router at the same budget reaches 66.2% (Fig. 3). The gate captures about 53% of the accuracy gain for 30% of the retrieval spend, the oracle about 71%; the difference is the ceiling on what learned routing is worth.

Table 4 is not a claim about any real system. The mechanism it conveys is that the same restriction share and the same retrieval accuracy can, once weighted by quality and by cost, call for entirely different engineering decisions.

Table 4 | Summary of illustrative calculations (fixed random seed 20261104)

Calculation	Key settings	Result	Equation
Quality-weighted restriction	20,000 sources; restriction probability rises with quality	ρ_{tok} 11.3%, ρ_q 16.9%, top decile 43.0%	Eqs (1)–(2)
Effective price per retrieval	3,000 back-filled tokens at US$0.30 per million	US$0.00165 (low tier), US$0.0109 (high tier)	Eq. (4)
Break-even query volume	US$20,000 fixed, US$2,000 per refresh, 3 retrievals per query	$\tau = 1$ month: 8.89M / 1.35M; $\tau = 12$ months: 4.44M / 0.67M	Eq. (6)
Budgeted gating	10,000 queries, 8% retrieval harm, 30% budget	Accuracy 58.0%→64.1% (oracle 66.2%, always 69.4%); fees US$32.7→9.8 per 1,000	Eq. (5)

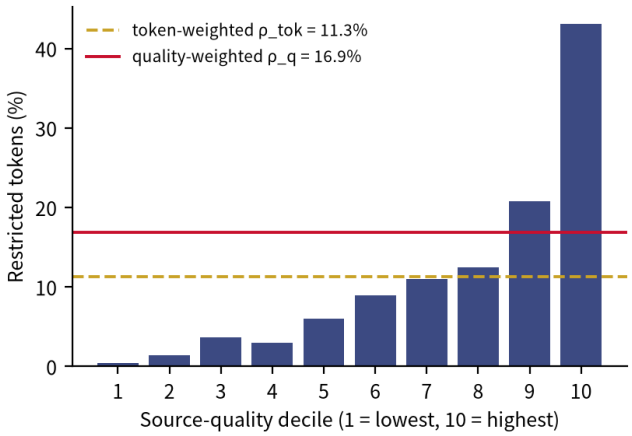


Fig. 1 | Restriction rises with source quality. The token-weighted restriction rate (dashed) understates the quality-weighted rate (solid); the gap equals the covariance term in equation (2). Illustrative data, fixed random seed, computed by the HX Research Team.

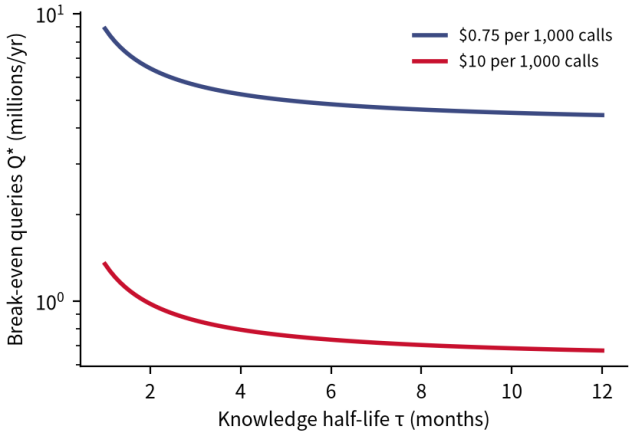


Fig. 2 | Break-even query volume for internalizing versus retrieving. Above a curve, internalizing is cheaper; below it, retrieving is cheaper; higher retrieval prices and longer knowledge half-lives lower the break-even point. Illustrative parameters, computed by the HX Research Team from equation (6).

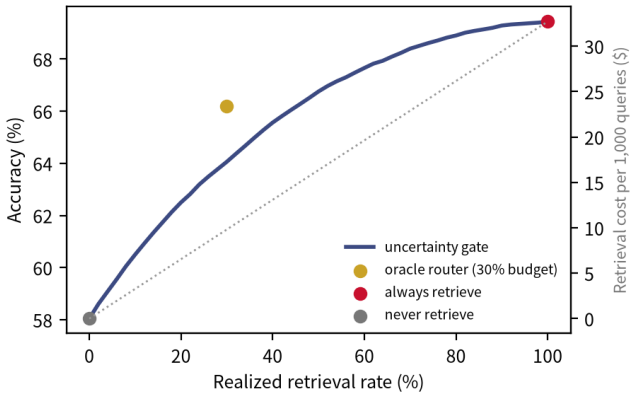


Fig. 3 | Accuracy versus retrieval rate under budgeted gating. The uncertainty gate yields most of its gain at low retrieval rates; the oracle router marks the ceiling for the same budget. Illustrative data, fixed random seed, computed by the HX Research Team.

Forecasts and research agenda

The following are the views of the HX Research Team, not statements of fact.

The research agenda contains four open problems. First, how can ρ_q be estimated without accessing blocked content? The quality score w_d itself requires the content — a self-referential problem. Second, can data attribution support payment by contribution? Influence functions now scale to models with 52B parameters[75], but turning them into a settleable unit of account remains distant. Third, can the causes of retrieval harm be predicted, turning $\Delta(x)$ in equation (5) from an after-the-fact statistic into a prior estimate? Fourth, when training and retrieval are declared separately and what a model "remembers" conflicts with what it "finds", which should prevail, and how should the conflict be shown to users?

Table 5 | Five forecasts (views)

No.	Forecast	Horizon	Main evidence	Falsified if	Enterprise implication
1	Technical reports and data summaries of major models will routinely disclose how training-preference signals were handled and the corpus snapshot date; "restriction by quality quantile" will start to appear on data cards	12–24 months	Shared preference signal[16,17]; EU Article 53 training-content summary[46]; measured quality skew[13,14]	By end-2028, technical reports of major open-weight models still say nothing about preference-signal handling	Add preference handling and snapshot dates to due-diligence checklists for built or procured models
2	Agent evaluations will report retrieval calls and fees per solved task alongside accuracy, and fixed-corpus evaluation will become standard for search agents	6–18 months	Metered retrieval[22,23]; fixed-corpus benchmark[74]; budget-aware evaluation[36]	By end-2027, leaderboards of the main search-agent benchmarks still report accuracy only	Add call counts and p_{eff} to internal acceptance criteria
3	In public pretraining recipes, rewritten and synthetic data will commonly exceed a fifth of the mixture, written in as appended rather than substituted data	12–24 months	Rewriting scaling law[32]; accumulation avoids collapse[59]; data-constrained prescriptions[53]	By 2028, rewrite shares in major public recipes remain below 10% with no visible quality loss	Log generator versions and rewrite shares in the recipe ledger
4	Open, cross-vendor standards drafts for pricing and receipts in machine access will appear, moving HTTP 402-style negotiation from proprietary CDN features to standards	12–24 months	Pay-per-crawl[25]; terms.txt[26]; pricing agents[27]; limits of the AIPREF charter[21]	By end-2028, no standards body has adopted a work item on payment and receipts for machine access	Budget and procure for sources that are "searchable but paid"
5	For stable internal knowledge, enterprises will increasingly adopt private corpora with small domain models and in-house retrieval, setting the scope of internalization by the break-even in equation (6)	6–12 months	Retrieval-augmented scaling[33,43]; per-call billing[22]; gating[37,38]	Within a year, enterprise surveys show external retrieval fees below 5% of total inference spend, managed by no one	Keep an internalization list and refresh cadence for high-volume knowledge domains

Where common industry views may be wrong

"Only a small share of data is restricted." True by token count, false by quality weight. Equation (2) shows that the gap equals the covariance of quality and restriction divided by mean quality, and current measurements show that covariance to be positive[11,13,14]. Budgeting by token share underestimates the cost of replacing data.

"Synthetic data can replace real data indefinitely." Synthetic rewriting does follow scaling laws[32], but substitutive self-consuming loops collapse[57,58], and verifiers can only pull a model towards their own knowledge centre[60]. Synthetic data amplify the information that remains, not the information unique to blocked sources.

"Retrieval always beats no retrieval." Retrieval harm is non-negligible, and thresholds may miss target usage on new data[36]; in long contexts, evidence position affects how well it is used[35]. Retrieval should be gated, not on by default.

"Tool calls are cheap, so retrieval cost is negligible." Per-call prices can differ by an order of magnitude across providers[22,23], and back-filled tokens are billed on top[24]; when agents run continuously, call counts grow with uptime[28,29]. In our illustrative calculation, back-filled tokens already cost more than the call itself at the lower price point.

"Preference signals are binding, so compliance makes us safe." The robots.txt specification states plainly that its rules are not access authorization[34], and the AIPREF charter excludes technical enforcement[21]; the legal basis for compliance still comes from copyright law and sector regulation[46,48]. Preference signals lower the cost of expression; they do not replace legal judgement.

"Openly licensed corpora cannot train good models." Common Pile's 7B models perform competitively with Llama 1 and Llama 2 7B at similar compute (per the paper)[62], and The Stack showed that permissively licensed code alone can match previously reported code-benchmark results (per the paper)[63]. The weakness of openly licensed corpora is coverage and freshness, not trainability.

Limitations

This paper has five limitations. First, quality-weighted restriction depends on the definition of the quality score w_d; different classifiers will give different ρ_q, and we argue only for the direction and structure of the gap, not for its value on real corpora. Second, the implementation details of preference signals are still changing, and some providers' timelines come from announcements and press reports[16,18]; forecasts based on them may need adjustment. Third, price lists can change at any time; the prices in our illustrative calculations serve only to show structure. Fourth, equation (6) assumes internalized knowledge is as good as retrieved knowledge, whereas internalizing can introduce hallucination or staleness and retrieval can introduce source noise. Fifth, nothing here is legal advice; specific copyright and data-compliance judgements should be made by qualified professionals.

Conclusion

Returning to the central thesis: the open web is turning from free feedstock into a metered input. On the training side, preference signals let "may this be trained on?" be declared once and apply everywhere, while restrictions concentrate on high-quality sources, so token-share losses are systematically understated. On the inference side, per-call billing turns every retrieval into an explicit cost item, amplified as agents run continuously. The marginal source of model capability is therefore shifting from crawl scale to two things that can — and must — be computed: rights-aware data mixtures and budgeted retrieval.

In the context of HX's products, the thesis lands in two places. According to the company website, Deep Hope S1 supports on-premises deployment, offers RAG over company policy, industry regulation and product material as its factual base, cites the source of every answer and declines to answer without one, and builds specialized capability through industry-model distillation[76] — route R5 in Table 3. The Harness governance engine of the Enterprise Agent Platform executes tool calls through a controlled MCP/API interface, with the governance layer holding the final say[77], which can carry the call ledger and budgeted gating of Algorithm 1. For enterprises, the transferable asset is not a corpus crawled once but a continuously updated ledger of rights and calls.

Key points:

1. Read data through equation (2): weight restriction by quality; the gap equals covariance over mean quality.
2. Fill the gap through equation (3): license first, repeat second, rewrite last — and append rewrites, never substitute them.
3. Price retrieval through equation (4): the effective price per retrieval includes back-filled tokens.
4. Govern calls through equation (5): gate retrieval, and report realized usage and harm rates.
5. Partition knowledge through equation (6): internalize high-volume, stable knowledge; retrieve the rest.

Data availability

This article produces no new experimental data. Cited figures come from the listed references, official documentation and the company website; illustrative calculations are generated by scripts under the stated assumptions.

Code availability

Scripts for the illustrative calculations and Figs. 1–3 are in src/figs.py (Python, fixed random seed 20261104), provided with this package for verification.

Competing interests

The authors are employees of Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. Deep Hope S1 and the Enterprise Agent Platform Harness mentioned in this article are public products of the company.

References

1. Common Crawl Foundation. Common Crawl: open repository of web crawl data. `https://commoncrawl.org/` (2026).
2. Raffel, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020); `https://arxiv.org/abs/1910.10683`.
3. Dodge, J. et al. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (2021); `https://arxiv.org/abs/2104.08758`.
4. Gao, L. et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. Preprint at `https://arxiv.org/abs/2101.00027` (2020).
5. Penedo, G. et al. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. Preprint at `https://arxiv.org/abs/2306.01116` (2023).
6. Soldaini, L. et al. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (2024); `https://arxiv.org/abs/2402.00159`.
7. Penedo, G. et al. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. Preprint at `https://arxiv.org/abs/2406.17557` (2024).
8. Li, J. et al. DataComp-LM: In search of the next generation of training sets for language models. Preprint at `https://arxiv.org/abs/2406.11794` (2024).
9. Kaplan, J. et al. Scaling Laws for Neural Language Models. Preprint at `https://arxiv.org/abs/2001.08361` (2020).
10. Hoffmann, J. et al. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems* **35** (2022); `https://arxiv.org/abs/2203.15556`.
11. Longpre, S. et al. Consent in Crisis: The Rapid Decline of the AI Data Commons. In *Advances in Neural Information Processing Systems* **37**, Datasets and Benchmarks Track (2024); `https://arxiv.org/abs/2407.14933`.
12. Longpre, S. et al. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence* **6**, 975–987 (2024); `https://doi.org/10.1038/s42256-024-00878-8`.
13. Steinacker-Olsztyn, N., Gosain, D. & Dao, H. Is Misinformation More Open? A Study of robots.txt Gatekeeping on the Web. In *Proc. ACM Web Conference 2026* (2026); `https://arxiv.org/abs/2510.10315`.
14. Bouchaud, P. & Ramaciotti, P. Web Crawler Restrictions, AI Training Datasets & Political Biases. Preprint at `https://arxiv.org/abs/2510.09031` (2025).
15. Liu, E. et al. Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. Preprint

at https://arxiv.org/abs/2411.15091 (2024).
16. OpenAI. Making it easier for websites to express AI training preferences. https://openai.com/index/helping-websites-express-ai-training-preferences (2026).
17. Mueller, J. Supporting a shared AI training control for websites. Google Search Central Blog, https://developers.google.com/search/blog/2026/10/ai-training-control (2026).
18. Roth, E. The biggest names in AI are backing a new way to opt out of scraping. *The Verge*, https://www.theverge.com/tech/1023763/openai-google-microsoft-ai-scraping-standard-opt-out (2026).
19. Keller, P. & Thomson, M. A Vocabulary For Expressing AI Usage Preferences. Internet-Draft draft-ietf-aipref-vocab-08, IETF, https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab/ (2026).
20. IETF AI Preferences Working Group. Associating AI Usage Preferences with Content in HTTP. Internet-Draft draft-ietf-aipref-attach-05, https://datatracker.ietf.org/doc/draft-ietf-aipref-attach/ (2026).
21. IETF. AI Preferences (aipref) working group charter. https://datatracker.ietf.org/wg/aipref/about/ (2026).
22. DeepSeek. Search & Fetch (API documentation). https://api-docs.deepseek.com/guides/search/ (2026).
23. Anthropic. Web search tool (Claude Platform documentation). https://docs.anthropic.com/en/docs/agents-and-tools/tool-use/web-search-tool (2026).
24. OpenAI. API pricing (tools: web search). https://platform.openai.com/docs/pricing (2026).
25. Allen, W. & Newton, S. Introducing pay per crawl: Enabling content owners to charge AI crawlers for access. Cloudflare Blog, https://blog.cloudflare.com/introducing-pay-per-crawl/ (2025).
26. Chowdhury, R. terms.txt: A Consent and Compensation Protocol for Agentic Web Access. Preprint at https://arxiv.org/abs/2609.11152 (2026).
27. Archer, R., Ghili, S. & Haghpanah, N. Pay-Per-Crawl Pricing for AI: The LM-Tree Agent. Preprint at https://arxiv.org/abs/2604.01416 (2026).
28. DeepSeek Harness. Exploring Super-agent Design: Towards a More Autonomous Agent System. X (formerly Twitter), https://x.com/DeepSeekHarness/status/2116857257245987236 (2026).
29. Chang, C. et al. From LLM Inference to Agentic Workloads: Characterization and Implications for Serving Systems. Preprint at https://arxiv.org/abs/2608.15127 (2026).
30. Villalobos, P. et al. Will we run out of data? Limits of LLM scaling based on human-generated data. Preprint at https://arxiv.org/abs/2211.04325 (2022).
31. Muennighoff, N. et al. Scaling Data-Constrained Language Models. In *Advances in Neural Information Processing Systems* **36** (2023); https://arxiv.org/abs/2305.16264.
32. Roberts, N. et al. Rewriting Scaling Laws for Synthetic Data with Compute and Data Constraints. Preprint at https://arxiv.org/abs/2610.31867 (2026).
33. Shao, R. et al. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore. In *Advances in Neural Information Processing Systems* **37** (2024); https://arxiv.org/abs/2407.12854.
34. Koster, M., Illyes, G., Zeller, H. & Sassman, L. Robots Exclusion Protocol. RFC 9309, IETF, https://www.rfc-editor.org/rfc/rfc9309.html (2022).
35. Liu, N. F. et al. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157–173 (2024); https://arxiv.org/abs/2307.03172.
36. Qian, P. et al. When Should Active RAG Retrieve? A Budget-Aware Evaluation of Utility, Calibration, and Cost. Preprint at https://arxiv.org/abs/2607.24010 (2026).
37. Wang, Y., wei, L. & Ling, H. Retrieval as a Decision: Training-Free Adaptive Gating for Efficient RAG. Preprint at https://arxiv.org/abs/2511.09803 (2025).
38. Lin, T.-H. et al. AdaSearch: Balancing Parametric Knowledge and Search in Large Language Models via Reinforcement Learning. Preprint at https://arxiv.org/abs/2512.16883 (2025).
39. Huang, Z., Yuan, X., Ju, Y., Zhao, J. & Liu, K. Reinforced Internal-External Knowledge Synergistic Reasoning for Efficient Adaptive Search Agent. Preprint at https://arxiv.org/abs/2505.07596 (2025).
40. Luo, Q.-W. et al. Heterogeneous Co-Evolution Learning for Tool-Efficient Search Agents with Zero Labels. Preprint at https://arxiv.org/abs/2610.39682 (2026).
41. Asai, A., Wu, Z., Wang, Y., Sil, A. & Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proc. International Conference on Learning Representations* (2024); https://arxiv.org/abs/2310.11511.
42. Jeong, S., Baek, J., Cho, S., Hwang, S. J. & Park, J. C. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics* (2024); https://arxiv.org/abs/2403.14403.
43. Borgeaud, S. et al. Improving language models by retrieving from trillions of tokens. In *Proc. 39th International Conference on Machine Learning* (PMLR, 2022); https://arxiv.org/abs/2112.04426.
44. Izacard, G. et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *J. Mach. Learn. Res.* **24**, 1–43 (2023); https://arxiv.org/abs/2208.03299.
45. W3C TDMRep Community Group. TDM Reservation Protocol (TDMRep), Final Community Group Report. https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/ (2024).
46. European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 53: Obligations for providers of general-purpose AI models. https://artificialintelligenceact.eu/article/53/ (2024).
47. European Commission. The General-Purpose AI Code of Practice. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (2025).
48. 国家互联网信息办公室等. 生成式人工智能服务管理暂行办法（Interim Measures for the Management of Generative AI Services）. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm (2023).
49. Cheng, J. et al. Dated Data: Tracing Knowledge Cutoffs in Large Language Models. Preprint at https://arxiv.org/abs/2403.12958 (2024).
50. Carlini, N. et al. Extracting Training Data from Large Language Models. In *Proc. 30th USENIX Security Symposium* (2021); https://arxiv.org/abs/2012.07805.
51. Carlini, N. et al. Quantifying Memorization Across Neural Language Models. In *Proc. International Conference on Learning Representations* (2023); https://arxiv.org/abs/2202.07646.
52. Cooper, A. F. et al. Extracting memorized pieces of (copyrighted) books from open-weight language models. Preprint at https://arxiv.org/abs/2505.12546 (2025).
53. Baron, E. et al. Improving Data-Constrained Scaling with Hyperparameter Transfer and Ensembling. Preprint at https://arxiv.org/abs/2610.37842 (2026).
54. WANG, Z. & Kanamori, T. Selecting Repetition Counts Across Model Scales in Data-Constrained Pretraining. Preprint at https://arxiv.org/abs/2610.05126 (2026).
55. Prabhudesai, M., Wu, M., Zadeh, A., Fragkiadaki, K. & Pathak, D. Diffusion Beats Autoregressive in Data-Constrained Settings. Preprint at https://arxiv.org/abs/2507.15857 (2025).
56. Gunasekar, S. et al. Textbooks Are All You Need. Preprint at https://arxiv.org/abs/2306.11644 (2023).
57. Shumailov, I. et al. AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024); https://doi.org/10.1038/s41586-024-07566-y.
58. Alemohammad, S. et al. Self-Consuming Generative Models Go MAD. In *Proc. International Conference on Learning Representations*

(2024); https://arxiv.org/abs/2307.01850.
59. Gerstgrasser, M. et al. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. Preprint at https://arxiv.org/abs/2404.01413 (2024).
60. Yi, B., Liu, Q., Cheng, Y. & Xu, H. Escaping Model Collapse via Synthetic Data Verification: Near-term Improvements and Long-term Convergence. Preprint at https://arxiv.org/abs/2510.16657 (2025).
61. Baack, S. et al. Towards Best Practices for Open Datasets for LLM Training. Preprint at https://arxiv.org/abs/2501.08365 (2025).
62. Kandpal, N. et al. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. Preprint at https://arxiv.org/abs/2506.05209 (2025).
63. Kocetkov, D. et al. The Stack: 3 TB of permissively licensed source code. Preprint at https://arxiv.org/abs/2211.15533 (2022).
64. Lozhkov, A. et al. StarCoder 2 and The Stack v2: The Next Generation. Preprint at https://arxiv.org/abs/2402.19173 (2024).
65. Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems* **33** (2020); https://arxiv.org/abs/2005.11401.
66. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M.-W. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proc. 37th International Conference on Machine Learning* (PMLR, 2020); https://arxiv.org/abs/2002.08909.
67. Nakano, R. et al. WebGPT: Browser-assisted question-answering with human feedback. Preprint at https://arxiv.org/abs/2112.09332 (2021).
68. Schick, T. et al. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems* **36** (2023); https://arxiv.org/abs/2302.04761.
69. Yao, S. et al. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proc. International Conference on Learning Representations* (2023); https://arxiv.org/abs/2210.03629.
70. Jin, B. et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. Preprint at https://arxiv.org/abs/2503.09516 (2025).
71. Wei, J. et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. Preprint at https://arxiv.org/abs/2504.12516 (2025).
72. Grossman, R. et al. How Generative AI Disrupts Search: An Empirical Study of Google Search, Gemini, and AI Overviews. Preprint at https://arxiv.org/abs/2604.27790 (2026).
73. Noroozian, A. et al. Generative AI and the Future of the Digital Commons: Five Open Questions and Knowledge Gaps. Preprint at https://arxiv.org/abs/2508.06470 (2025).
74. Chen, Z. et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. Preprint at https://arxiv.org/abs/2508.06600 (2025).
75. Grosse, R. et al. Studying Large Language Model Generalization with Influence Functions. Preprint at https://arxiv.org/abs/2308.03296 (2023).
76. 弘犀智能（Great Unicorn Intelligence）. Deep Hope S1 · 推训一体大模型平台（Deep Hope S1 · unified inference–training LLM platform）. https://www.hxallai.com/platform/deep-hope/ (2026).
77. 弘犀智能（Great Unicorn Intelligence）. 企业智能体平台 · Harness（Enterprise Agent Platform · Harness）. https://www.hxallai.com/platform/agent/ (2026).

Copyright Notice: © 2026 Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. All rights reserved. Any reposting, quotation in commentary, or adaptation must credit the authors (Great Unicorn Intelligence Research Team) and cite the source (Great Unicorn Intelligence, https://www.hxallai.com).