UNDERSTANDING AI · WRITTEN EDITION · 2026-09-11
Day 7 · typeset 2026-09-11 08:24Z

Since August 2025, companies placing general-purpose AI models on the European market have been required to publish a summary of the content used to train them, and since 2 August 2026 the European Commission has been able to fine those that do not. The documents exist, they are dated, and anyone can download them. Read closely, they describe less a corpus than a filing cabinet. Six decades of corpus-building show how much of the selection of training text has moved into rules.

1. A deadline, and what arrived before it

Article 53(1)(d) of Regulation (EU) 2024/1689 requires a provider of a general-purpose AI model to publish "a sufficiently detailed summary about the content used for training … according to a template provided by the AI Office". The European Commission published that template on 24 July 2025, and the obligation began to apply on 2 August 2025 to models placed on the market from that date; under Article 111(3), providers of models already on sale have until 2 August 2027.

The power to enforce it arrived a year later, and the statute is precise about the gap. Article 113 provides that Chapter V — the general-purpose-AI chapter, which contains the summary obligation — applies from 2 August 2025 "with the exception of Article 101". Article 101 is headed *Fines for providers of general-purpose AI models*, and permits the Commission to impose penalties, for intentional or negligent infringement, of up to 3% of worldwide annual turnover or €15 million, whichever is higher. It took effect on the Regulation's general date of application: **2 August 2026**.

Publication clusters around the second date rather than the first; nothing in the public record shows that one caused the other.

Europe's is not the only such duty. California's AB 2013, approved on 28 September 2024 as Chapter 817 of that year's statutes and now section 3111 of the state's Civil Code, requires developers of generative AI systems or services released since January 2022 and made available to Californians to post, "on or before January 1, 2026" and before later covered releases, documentation that includes a high-level summary of their training datasets, beginning with "the sources or owners of the datasets". A federal district court refused xAI a preliminary injunction against it on 4 March 2026, and xAI's appeal is before the Ninth Circuit. OpenAI's page under the Californian statute, in an Internet Archive capture of 21 January 2026, is one company-wide document describing "datasets containing trillions of tokens" and naming none of them.

Figure 1. Public training-content summaries, before and after the enforcement date

5
summaries found by 12 January 2026
the paper's abstract; its body counts four, plus a non-summary file, as of 20 January

44
entries on the same team's tracker, unchanged 1–11 September 2026
19 further models judged, on preliminary analysis, to need one

2 Aug 2026
date the power to fine took effect
the obligation itself began 2 August 2025

Counts: Blankvoort, Pandit and Gahntz, arXiv 2603.13270 v1 (26 February 2026) for the January figure; their public tracker at aial.ie, read on 1 September 2026 and again on 11 September 2026, when the page was byte-for-byte unchanged and reported a last modification on 31 August. Dates: Regulation (EU) 2024/1689, Articles 53, 101 and 113.

▼ **Table view**

Figure 1. Public training-content summaries, before and after the enforcement date

Measure	Value
summaries found by 12 January 2026	5
entries on the same team's tracker, unchanged 1–11 September 2026	44
date the power to fine took effect	2 Aug 2026

In January 2026 three researchers — Dick Blankvoort and Harshvardhan Pandit of the AI Accountability Lab at Trinity College Dublin, and Maximilian Gahntz — assessed what their paper calls the five public summaries "found through an exhaustive search process". Four came from Hugging Face (for SmolLM3), the Swiss AI Initiative (for Apertus), the Polish open-source collaboration SpeakLeash (for Bielik) and the Israeli image-model company Bria. For the fifth, Microsoft's Phi-4, the authors record that they "did not find an explicitly published summary", but found a document in the model's repository whose structure was similar to the template and which, in their words, "was not explicitly mentioned as a public summary".

The large providers published documents dated between late June and early August 2026: OpenAI's for GPT-5.5 on 26 June (revised on 29 July), Google's for Gemini on 2 July, xAI's for Grok 4.5 on 8 July, Inkling's on 15 July, Anthropic's for Claude Opus 5 and OpenAI's for GPT-5.6 Luna on 23 July, and Meta's for Muse Spark on 4 August. Each provider publishes its summary on its own website. By its own summary's account, GPT-5.5 had been on the Union market since 23 April 2026, two months before the first version of that summary. OpenAI's summary for GPT-6 Astra carries a last-update date of 2 September 2026 and gives 3 September as the model's date of placement — a summary dated before its model's launch rather than after it. The chronology is a matter of record; why each provider published when it did is not visible from outside.

The same team keeps a public tracker. On 1 September 2026 it listed 44 entries — 39 graded and five under evaluation, counting model variants separately and including the Phi-4 file — and 19 further models that the researchers, "based on our preliminary analysis", judge to need a summary and could not find one. Fetched again on 11 September, the page had not changed by a single byte; its server reports a last modification on 31 August, so the GPT-6 Astra summary is not on it. The count records the tracker's own pace, not the providers'.

2. Two misreadings, and what a compelled disclosure contains

Two readings of these summaries are common, and the evidence supports neither. The first is that the public now knows what the models were trained on. The second is its mirror image: that the exercise is theatre, the companies took the internet wholesale, and nobody chose anything. The first fails on the documents themselves. The second fails on six decades of practice: somebody has always chosen what goes into a corpus, and what has changed is how much of the choosing is done by rules.

The template is a form with tick-boxes. Its largest volume option for text is "More than 10 trillion tokens", which is open-ended: it is ticked by OpenAI, Anthropic, xAI and by a 70-billion-parameter academic model alike, so across that range the tick-box alone distinguishes nothing.

One field asks for a "Summary of the most relevant domain names crawled". Six summaries answer it as follows.

Provider (model)	Document's own date	Answer to "most relevant domain names crawled"	Individual websites named
OpenAI (GPT-5.5)	29 Jul 2026 (first version 26 Jun)	academic, research, patent and other technical repositories; legal and government resources; document-hosting and sharing services; community and general-interest sites; region-specific portals	0
xAI (Grok 4.5)	8 Jul 2026	academic/research repositories, patent and technical databases, legal/government resources, document-sharing platforms, community sites, region-specific portals	0
Anthropic (Claude Opus 5)	23 Jul 2026	technical documentation, open-source software, reference sites, document sharing sites, math sites; *"Top-level domains such as .com, .org, and .net are included"*	0
Google (Gemini 3 Pro family)	2 Jul 2026	publicly available websites across educational, government, legal and research sectors; a hyperlink to a list of Google's crawlers	0
Inkling	15 Jul 2026	resources and repositories spanning academic, scientific, mathematical, code-related and general-purpose content	0
Meta (Muse Spark)	4 Aug 2026	a pointer to the company's developer centre	0

The first two answers are near-identical in content and in order. Two documents from competing companies converge on the same categories in the same sequence; how that came about is not visible from outside. OpenAI's GPT-6 Astra summary of 2 September repeats the same list.

Where the same form asks for a list of large publicly available datasets, OpenAI's entire answer for GPT-5.5 is one sentence:

> The training data for GPT-5.5 includes text from Common Crawl.

That is the whole of that field. Elsewhere on the same page the summary adds an exclusion rule:

> OpenAI also uses the U.S. Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list as a signal when deciding to exclude data from certain websites that have been recognized as persistently and repeatedly infringing copyright.

A list compiled for trade policy — naming marketplaces for counterfeit goods and pirated content — now serves as one input to what a language model is trained on. The summary does not say which edition of the list is used; what it records is the use, not the list.

3. The idea: a corpus is a built object

Much of what a deployed model appears to know at the moment it answers is text that ordinary software has placed in front of it: search results, files, earlier turns of a conversation. Beneath that run-time selection sits an older and far larger one — the text the model was trained on. Even the vocabulary through which a model reads is learned by counting over a large body of text, so whose text was in that body helps decide which languages are cheap to write in. A training corpus is not a natural object that happened to be lying around. It is built, and its builders' choices are recorded, when they are recorded at all, in documents like those above.

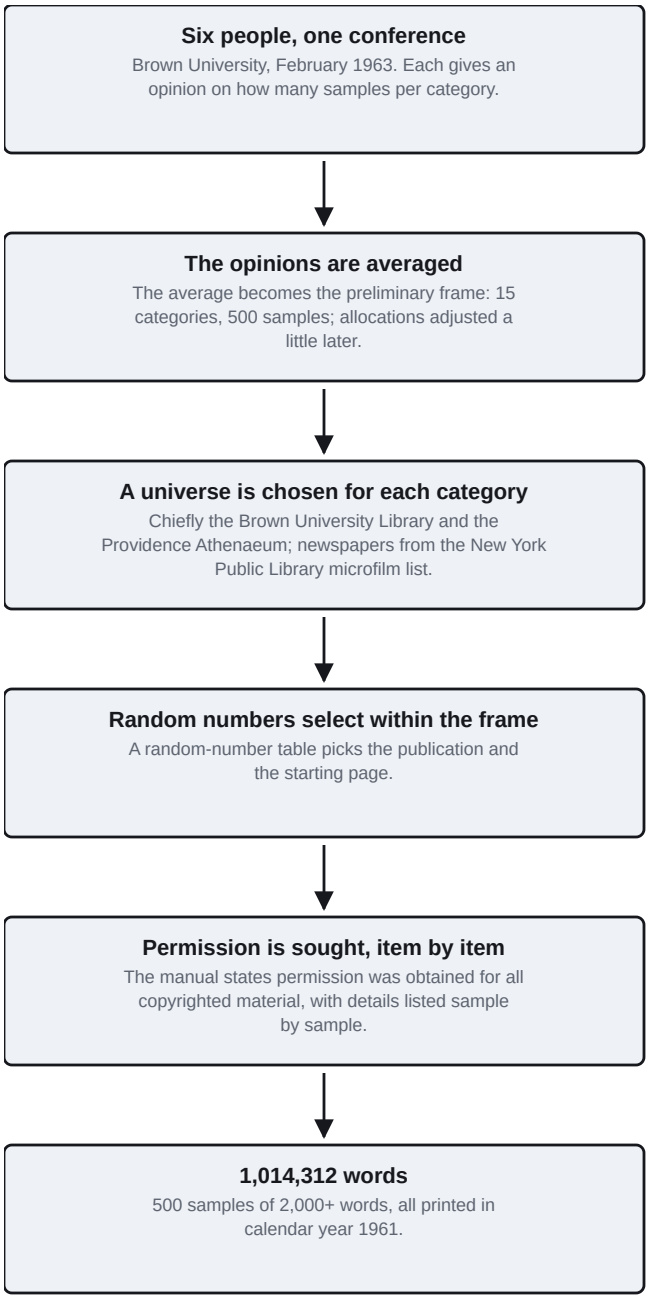
A corpus is a body of text collected on purpose. A classic early example is the Brown Corpus: 1,014,312 words of edited American prose, all of it printed in the calendar year 1961, divided into 500 samples of roughly two thousand words each.

Somebody had to decide what a million words of English *is*. The manual that accompanies the corpus, by W. Nelson Francis and Henry Kučera — first issued in 1964; the text available online is its 1979 revision — records who and how: the list of categories "was drawn up at a conference held at Brown University in February 1963", attended by John B. Carroll, W. Nelson Francis, Philip B. Gove, Henry Kučera, Patricia O'Connor and Randolph Quirk. The manual then says what those six people did:

> These figures were averaged to obtain the preliminary set of figures used.

Six opinions, averaged, set the preliminary shares of newspaper reportage, fiction, religious writing and the rest; the manual adds that a few changes were later made on the basis of experience, and that finer subdivision followed the proportions actually published in 1961. Verse was excluded as a sampling category on the ground that it "presents special linguistic problems"; drama was excluded as "the imaginative recreation of spoken discourse"; fiction was admitted, but no sample could be more than half dialogue. Some of the hobby and popular-lore material was chosen from "one of the largest second-hand magazine stores in New York City". Only after all of that did randomness enter: within each category, samples were drawn chiefly using a table of random numbers, and the starting page was drawn the same way.

Figure 2. How a corpus was assembled in 1963



Source: the Francis and Kučera manual (1964; revised 1971; revised and amplified 1979), Internet Archive capture of the ICAME web edition, 13 March 2024. Randomness operates inside a frame that human judgement drew first.

▼ **Table view**

Figure 2. How a corpus was assembled in 1963 — stages

#	Stage	Note
1	Six people, one conference	Brown University, February 1963. Each gives an opinion on how many samples per category.
2	The opinions are averaged	The average becomes the preliminary frame: 15 categories, 500 samples; allocations adjusted a little later.

#	Stage	Note
3	A universe is chosen for each category	Chiefly the Brown University Library and the Providence Athenaeum; newspapers from the New York Public Library microfilm list.
4	Random numbers select within the frame	A random-number table picks the publication and the starting page.
5	Permission is sought, item by item	The manual states permission was obtained for all copyrighted material, with details listed sample by sample.
6	1,014,312 words	500 samples of 2,000+ words, all printed in calendar year 1961.

Figure 2. How a corpus was assembled in 1963 — connections

From	To	Label
Six people, one conference	The opinions are averaged	
The opinions are averaged	A universe is chosen for each category	
A universe is chosen for each category	Random numbers select within the frame	
Random numbers select within the frame	Permission is sought, item by item	
Permission is sought, item by item	1,014,312 words	

The manual's account of permission is a single sentence:

> For all copyrighted material used, the permission of the copyright holder has been obtained.

It continues that the details of copyright permission appear in the listing of the samples, on pages 33 to 176. Each sample's entry in the manual includes a "Copyright statement" field. The permissions carried conditions: the manual tells commercial publishers and other non-academic organisations that public use of the corpus requires permission from Brown's linguistics department, which may ask them to obtain written permission from the individual copyright holders. Sample A07 of the corpus is the New York Times.

4. What replaces asking

A million words can be cleared by hand. Ten trillion tokens cannot be cleared that way: at that size permission is not sought item by item, and no one could read the result. **At this scale, rules select much of the text**, though builders still choose datasets and, in some cases, buy or license material.

Permission at scale is expensive rather than impossible. *The Common Pile v0.1*, released on 5 June 2025, is an eight-terabyte collection of public-domain and openly licensed text drawn from 30 sources; its authors trained two 7-billion-parameter models on one and two trillion tokens of it and report performance competitive with models of similar computational budget trained on unlicensed text. Whether a collection of that kind can support a frontier model has not been shown.

The published descriptions specify the rules. In February 2019 OpenAI described the selection rule for GPT-2's corpus:

> we scraped all outbound links from Reddit, a social media platform, which received at least 3 karma

The rule selects the page a link points to, not the discussion around it; karma is, roughly, upvotes minus downvotes. Its authors immediately qualify what the signal measures:

> This can be thought of as a heuristic indicator for whether other users found the link interesting, educational, or just funny

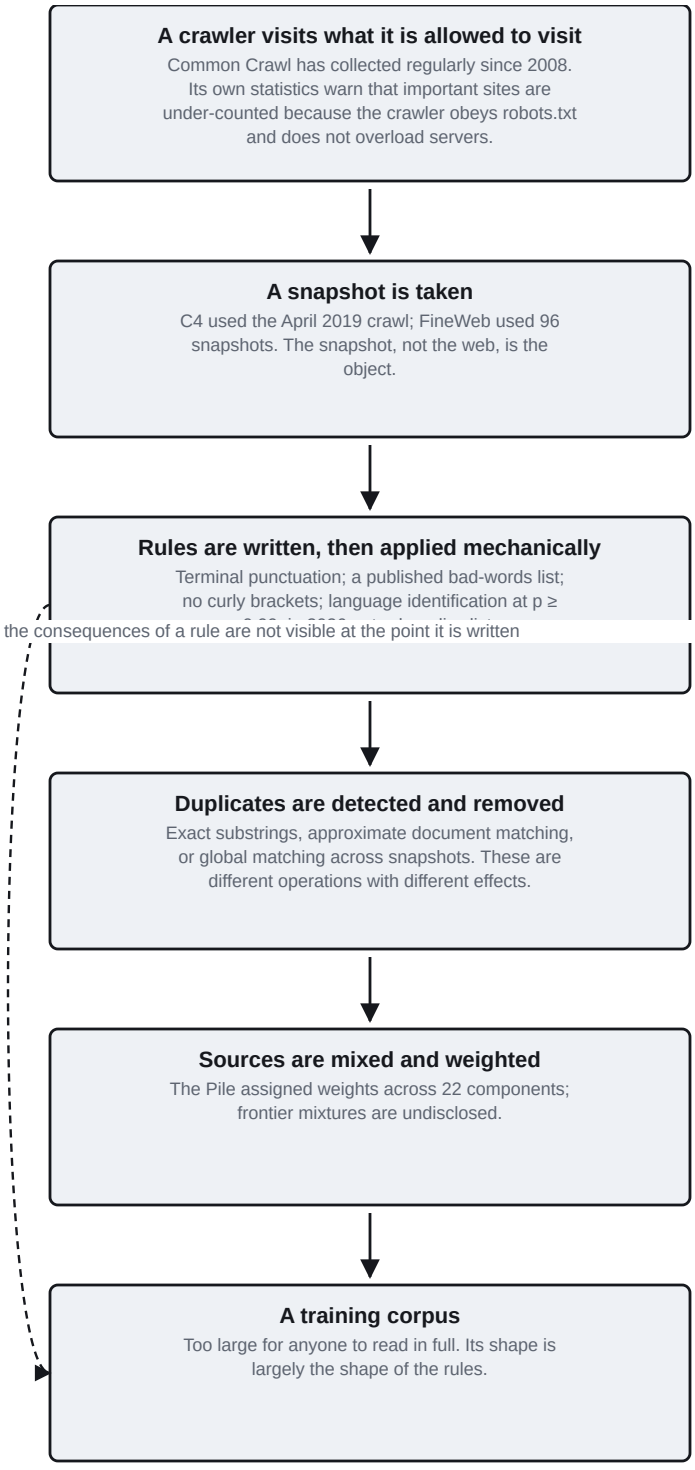
That corpus comprised 45 million links, reduced after deduplication and cleaning to "slightly over 8 million documents for a total of 40 GB of text". It excluded links created after December 2017, and the authors say they removed all Wikipedia articles to avoid overlap with their own evaluations.

Eight months later a Google team published the Colossal Clean Crawled Corpus, known as C4, built from the April 2019 Common Crawl snapshot under six cleaning rules and a language filter: retain only lines ending in a period, exclamation mark, question mark or end quotation mark; discard any page containing a term from the published "List of Dirty, Naughty, Obscene or Otherwise Bad Words"; drop any line containing the word Javascript; drop any page containing the phrase "lorem ipsum"; keep one copy of any three-sentence span that occurs more than once; and

> Since the curly bracket “{” appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.

Pages not classified as English with a probability of at least 0.99 were dropped as well. Later revisions of the same paper, in July 2020 and September 2023, list seven and then nine rules, and swap two thresholds between them — pages of fewer than five sentences and lines of fewer than three words in one, fewer than three sentences and five words in the other. The written description of the rules is itself a versioned document.

Figure 3. How a corpus is assembled now



Schematic of the crawl-based branch only; real corpora also draw on curated datasets, code, books and licensed material. Each stage is a decision made by people and tested on samples; none involves reading all of the material selected.

▼ **Table view**

Figure 3. How a corpus is assembled now — stages

#	Stage	Note

#	Stage	Note
1	A crawler visits what it is allowed to visit	Common Crawl has collected regularly since 2008. Its own statistics warn that important sites are under-counted because the crawler obeys robots.txt and does not overload servers.
2	A snapshot is taken	C4 used the April 2019 crawl; FineWeb used 96 snapshots. The snapshot, not the web, is the object.
3	Rules are written, then applied mechanically	Terminal punctuation; a published bad-words list; no curly brackets; language identification at p ≥ 0.99; in 2026, a trade-policy list.
4	Duplicates are detected and removed	Exact substrings, approximate document matching, or global matching across snapshots. These are different operations with different effects.
5	Sources are mixed and weighted	The Pile assigned weights across 22 components; frontier mixtures are undisclosed.
6	A training corpus	Too large for anyone to read in full. Its shape is largely the shape of the rules.

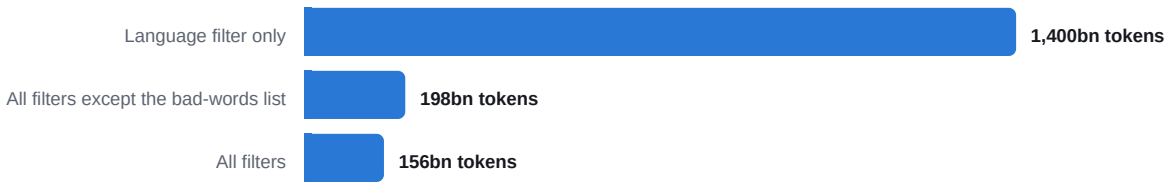
Figure 3. How a corpus is assembled now — connections

From	To	Label
A crawler visits what it is allowed to visit	A snapshot is taken	
A snapshot is taken	Rules are written, then applied mechanically	
Rules are written, then applied mechanically	Duplicates are detected and removed	
Duplicates are detected and removed	Sources are mixed and weighted	
Sources are mixed and weighted	A training corpus	
Rules are written, then applied mechanically	A training corpus	the consequences of a rule are not visible at the point it is written

5. What the rules did

Independent audits can measure effects that a builder's own description leaves unquantified. In April 2021 a team at the Allen Institute for AI and the University of Washington — Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld and Matt Gardner — documented C4, eighteen months after its publication, and released three versions of it: language-identified but otherwise unfiltered, filtered without the bad-words list, and fully filtered.

Figure 4. What each filtering stage removed



Dodge et al., Documenting the English Colossal Clean Crawled Corpus, arXiv 2104.08758 v1, 18 April 2021, Table 1. The bad-words list alone accounts for the fall from 198 to 156 billion tokens — about 21% of what reached it, computed from their figures.

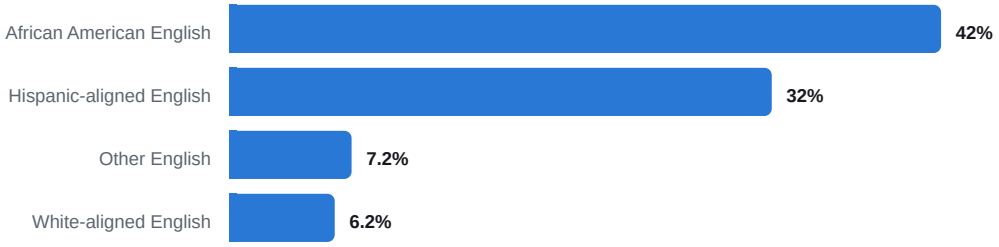
▼ Table view

Figure 4. What each filtering stage removed

Version of the corpus	Tokens
Language filter only	1,400bn tokens
All filters except the bad-words list	198bn tokens
All filters	156bn tokens

One published list of words removed about a fifth of the tokens that reached it — about 7.6% of the documents, which fell from 395 million to 365 million. It did not remove them evenly.

Figure 5. Share of documents removed by the bad-words filter, by dialect category



Same source. The categories are assignments made by a dialect-aware topic model trained on 60 million geolocated tweets, using each document's most likely dialect; they are not statements about authors, and the paper says identifying authors would be "infeasible and ethically questionable". Its own conclusion is that the findings suggest the blocklist disproportionately removes documents detected to be in dialects associated with minority identities.

▼ Table view

Figure 5. Share of documents removed by the bad-words filter, by dialect category

Dialect assigned by classifier	Removed
African American English	42%
Hispanic-aligned English	32%
Other English	7.2%
White-aligned English	6.2%

By the same audit's count, 97.8% of documents in the fully filtered corpus were assigned the white-aligned category, against 0.07% African American English.

The audit turned up a second surprise: the single most represented website in the cleaned corpus is `patents.google.com`, a substantial fraction of it machine-translated or produced by optical character recognition. The second is English Wikipedia.

Three months later a team led by Katherine Lee and Daphne Ippolito, with Nicholas Carlini among the co-authors, went looking for repetition. Lee was also one of C4's own authors, so this check came partly from inside. They found a single 61-word English passage occurring in the corpus **61,036 times verbatim in the training data and 61 times in the validation set**. Their footnote prints it. It ends:

> believe me, brilliant ideas would be perfect if it can be applied in real and make the people around you amazed!

It reads like filler from a wedding website. C4's published recipe included a rule keeping a single copy of any three-sentence span that recurs; the passage is in the released corpus 61,036 times regardless. By the authors' count those copies are 0.02% of the samples in each split, and because some sit in the validation set they create the overlap between training and evaluation data that, the authors warn, leads researchers to over-estimate accuracy.

6. Deduplication, and what it does not do

The remedy that paper proposed was far stricter deduplication — exact matching of long repeated substrings and approximate matching of near-identical documents — and its measured benefits are real: in the authors' experiments, models trained on deduplicated data "emit memorized text ten times less frequently" during unprompted generation, need fewer training steps for the same accuracy, and show no worse perplexity. The authors add that it can "reduce train-test overlap, which affects over 4% of the validation set of standard datasets".

It does not eliminate memorisation. An overlapping group of authors wrote seven months later:

> However, we find that memorization does still happen, even with just a few duplicates—thus, deduplication will not perfectly prevent leakage.

In the model families that 2022 work tested, larger models and more heavily duplicated examples yielded more extractable training text, and longer prompts exposed more of what a fixed model had memorised; the authors describe existing extraction measurements as lower bounds.

A further result points the opposite way from the intuition that more deduplication is always better. When Hugging Face built FineWeb in 2024, applying deduplication globally across 96 Common Crawl snapshots removed as much as 90% of the base data from the oldest snapshots — and the models trained on what remained showed little improvement over models trained on data that had not been deduplicated at all. The team then compared the 10% retained from an old snapshot with a separately deduplicated slice of the 90% it had discarded: models trained on the discarded material did **better**, and inspection found more

advertisements, keyword lists and badly formatted text in what had been kept. In their words, the result "challenged our initial assumption that global deduplication would inevitably result in higher benchmark scores."

"Deduplication" bundles three separate choices: what counts as a match (an exact passage or a near-identical document), where the search runs (within one crawl or across many), and which copy is kept. C4's three-sentence rule, the 2021 paper's exact-substring and near-duplicate matching and FineWeb's cross-snapshot procedure make different choices on each, and a result about one is not a result about the others.

7. Two fights, running in parallel

Through 2026 the question of training data has been contested in two venues that ask different things.

A courtroom asks whether the material could lawfully be used. In *Bartz v. Anthropic PBC* (N.D. Cal., 3:24-cv-05417), Judge Araceli Martínez-Olguín granted final approval on 20 July 2026 to a non-reversionary settlement fund of **$1.5 billion**, covering **482,460 works**, with an estimated per-work payment of approximately **$3,000** before costs and fees — which the order describes as "four times the minimum statutory damages amount for willful infringement". Three documents in that case are routinely conflated, and they establish different things:

Document	Date	What it establishes
Complaint	2024	The plaintiffs' allegations, without findings on their merits.
Dkt. 231, summary-judgment order (Judge Alsup)	23 Jun 2025	Findings on that record: that the company "pirated over seven million copies of books"; that using books to train was "exceedingly transformative" and a fair use; that building a permanent library of pirated copies was not itself a fair use
Dkt. 680, final approval	20 Jul 2026	That a settlement is fair, reasonable and adequate under Rule 23. It is not a finding of infringement, and no jury reached a verdict

The certified class covers owners of reproduction rights in books on the Works List — books "in the versions of LibGen or PiLiMi downloaded by Anthropic" that meet the order's ISBN, ASIN and copyright-registration criteria. The order records, without adopting it as a finding, the company's statement that neither of the named shadow-library datasets, nor any portion of them, was in the training corpus of any of its commercially released models.

The argument continues in other cases. On 1 September 2026 the United States government filed a statement of interest in the consolidated copyright cases against OpenAI in New York, arguing that training AI models on copyrighted material, "in and of itself", does not violate copyright law. It states the government's legal position, not a court's ruling.

A regulator asks whether the public will be told what the material was. The answers to its disclosure form diverge sharply.

Field	OpenAI, GPT-5.5	Apertus (ETH Zurich / EPFL)
Text volume	☒ More than 10 trillion tokens	☒ More than 10 trillion tokens — and, written in: 15 trillion tokens

Field	OpenAI, GPT-5.5	Apertus (ETH Zurich / EPFL)
Large public datasets (the same field)	one: Common Crawl	nine, each by URL, three with version numbers; a further field adds four more URLs and Wikipedia
Collection period, as each form scopes it	crawler collection: "Approximately 2018 – December 2025"	"from 2013 onward to a knowledge cutoff of March 2024 (CC-MAIN-2024-10)" — a named snapshot
Opt-outs	robots.txt signals for GPTBot, where available for listed domains	opt-outs recorded in January 2025 applied "retroactively in all earlier crawls since 2013"
Reproducibility	not addressed	filtering scripts published, with repository links

The retroactive removal has no counterpart in the other summaries: data from sites that, as of January 2025, had opted out of at least one common AI crawler was removed from crawls reaching back to 2013.

The comparison flatters neither side. The Swiss summary answers **No** to crawlers of its own, **No** to commercial licensing agreements, **No** to both user-data questions, **No** to synthetic data created by or for it and **No** to other sources. The likelier reading is that the corpus can be enumerated because it is built from things that were already public lists. No frontier provider has published an inventory of comparable detail.

That the frontier does not disclose is itself long documented and dated. OpenAI's GPT-4 technical report of 15 March 2023 states that "this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar", citing competitive and safety considerations. Google's Gemini 3 Pro model card names six classes of source and several processing techniques, including honouring robots.txt, without an inventory or a proportion. And an independent index of developer transparency, published in late 2025 by researchers at Stanford, Berkeley, Princeton and MIT, reported an average score of 40.69 out of 100 in its 2025 edition against 58 in 2024 — across company sets that differ between editions — with training data and training compute the areas of greatest opacity.

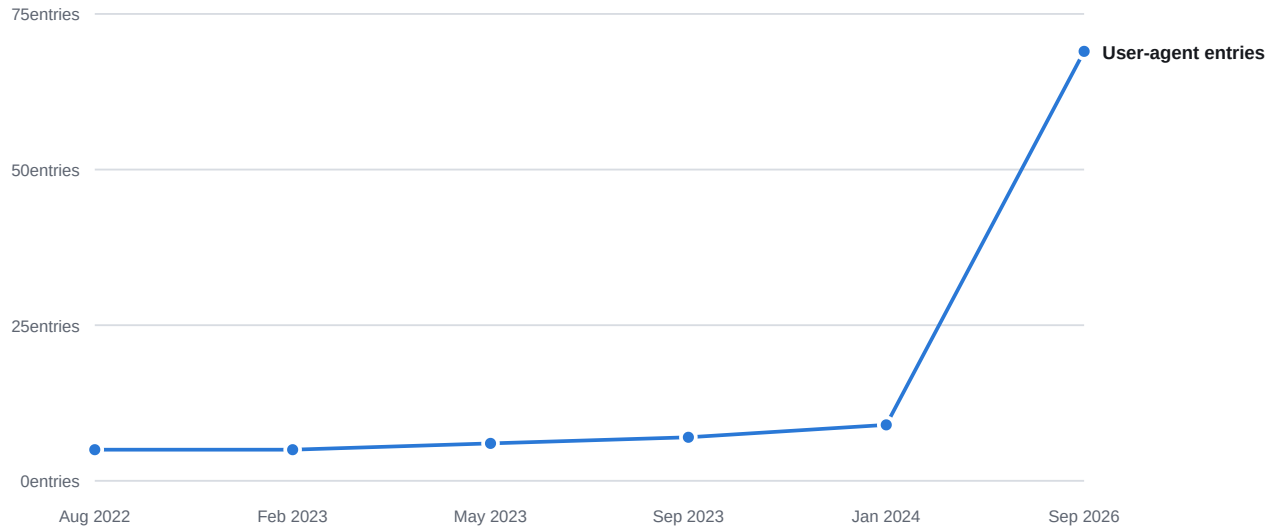
8. What is checkable, and when

The public tracker, crawler instructions and disclosure revisions are dated records against which later changes can be checked.

The tracker. Forty-four entries on 1 September 2026, against five summaries found by 12 January; nineteen models the same researchers judge to need a summary and for which they could not find one. On 11 September the page was byte-for-byte unchanged and did not list OpenAI's GPT-6 Astra summary of 2 September. Its next revision will show whether it has been added.

Two robots files. The text file a site uses to tell crawlers where not to go is the nearest thing to a dated public record of what a site asks of them, though what it says can depend on who asks. The version Reddit served on 11 September to ordinary requests from one machine is 538 bytes and ends with a wildcard and `Disallow: /`, above which its comments point to a content policy. Requests from the same machine presenting the user-agent strings of Google's and OpenAI's crawlers were refused outright, so these tests do not establish what a genuine crawler receives. The New York Times' file has grown differently.

Figure 6. Crawler directives in nytimes.com/robots.txt



Internet Archive captures of 1 August 2022, 1 February, 1 May and 1 September 2023 and 1 January 2024, plus live fetches on 1 and 11 September 2026. CCBot — Common Crawl's crawler — appears between the February and May 2023 captures; GPTBot by September 2023. Of the 69 user-agent entries in the current file one is the wildcard, 60 are told Disallow: / and nothing else, and one more is disallowed everywhere except two sections (the file has 56 Disallow: / lines, several shared by more than one crawler). One of the five August 2022 entries was also the wildcard. Between the two September fetches the file grew by two Allow lines for sports-discussion pages; the counts did not change.

▼ **Table view**

Figure 6. Crawler directives in nytimes.com/robots.txt

Capture	User-agent entries
Aug 2022	5entries
Feb 2023	5entries
May 2023	6entries
Sep 2023	7entries
Jan 2024	9entries
Sep 2026	69entries

Of those sixty-nine entries, sixty are told to stay away entirely, and a sixty-first is excluded from everything but two sections. One of them is CCBot, which means the newspaper that supplied sample A07 of the Brown Corpus — with permission recorded in the manual — now asks the crawler whose archive underpins C4, FineWeb and, by OpenAI's own summary, GPT-5.5 to stay away. The file is a request about future crawling, not a lock: it binds only crawlers that choose to obey it and removes nothing already collected.

Two copies of the same disclosure. OpenAI's public summary for GPT-5.5 is reachable at two addresses on the company's own content-delivery network, with different text and the same version label:

Copy	"Last update"	Status on 11 September 2026
the address the company's own help page links	29 July 2026	live; labelled "Version of the Summary: v1"

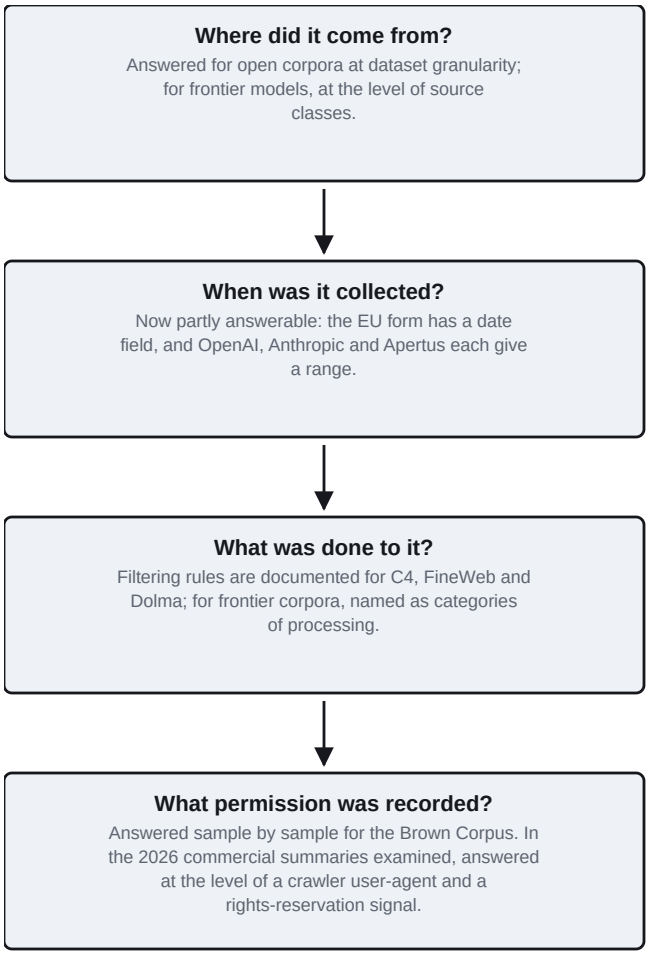
Copy	"Last update"	Status on 11 September 2026
an address carrying a document identifier	26 June 2026	live, still publicly downloadable; also labelled "v1"

Both files were byte-identical on 11 September to the copies fetched on 1 September. Compared sentence by sentence, the July text adds a statement that its disclosures on the types and quantities of training data apply to GPT-5.5 "and all subsequent releases in the model lifecycle", and drops a sentence about reducing the amount of personal data in training data that is largely duplicated elsewhere in the document. A copy of the June text, byte-identical to the one still online, sits in the university research group's archive under the date 14 July 2026, fifteen days before the revision's own date — independent evidence of the order in which the two appeared. Why the text changed, and whether the timing relative to the enforcement date signifies anything, is not visible from outside. What is visible is a legally mandated public record of what a model read, revised without a version increment.

9. Four questions of provenance

Provenance concerns where material came from, when it was collected, what was done to it and what permission was recorded — and who was responsible at each step.

Figure 7. The four questions, and where each is currently answerable



The four questions are stable; what has changed is the granularity at which each can be answered.

▼ **Table view**

Figure 7. The four questions, and where each is currently answerable — stages

#	Stage	Note
1	Where did it come from?	Answered for open corpora at dataset granularity; for frontier models, at the level of source classes.
2	When was it collected?	Now partly answerable: the EU form has a date field, and OpenAI, Anthropic and Apertus each give a range.
3	What was done to it?	Filtering rules are documented for C4, FineWeb and Dolma; for frontier corpora, named as categories of processing.
4	What permission was recorded?	Answered sample by sample for the Brown Corpus. In the 2026 commercial summaries examined, answered at the level of a crawler user-agent and a rights-reservation signal.

Figure 7. The four questions, and where each is currently answerable — connections

From	To	Label
Where did it come from?	When was it collected?	
When was it collected?	What was done to it?	
What was done to it?	What permission was recorded?	

For the Brown Corpus's million words, the manual records permission sample by sample. In 2026, for more than ten trillion tokens, the nearest thing to a list of websites in any of six compelled summaries is: dot com, dot org, dot net.

The change is not that nobody chooses. Six people averaging opinions in a room in 1963, a karma score of three on a message board in 2019, a punctuation mark, a curly bracket, a trade-policy list in 2026 — each is a decision, made by somebody, about what a machine would read. What changed is how much of the choosing is done by rules that operate on more text than anyone will ever read.

Choosing a rule is not the same as knowing its result. The fifth of the tokens that one word list removed, and the rates — 42% of documents classified as African American English against 6.2% of those classified white-aligned — at which it removed them, were measured eighteen months after C4 was published, by a team that had not built it. The wedding passage present 61,036 times was found later still, by a team that included one of the corpus's own authors. People chose the sources and the rules; the resulting collection still had to be checked.

10. Common readings the record does not support

Common reading	What the record supports
The models were trained on "the whole internet".	A crawl is a collection made at particular times; a training corpus is a further selection from it, usually combined with other sources.
Naming Common Crawl names the training data.	It names an upstream source, not the snapshots, the transformations or the final mixture.
Publicly available means free to use.	Access, asserted permission and conditions of reuse are separate facts; even the Brown Corpus's permissions came with conditions.
Deduplication fixes memorisation.	In the reported experiments, deduplication reduced measured memorisation without eliminating leakage.
A "clean" corpus is a representative one.	Clean means it passed particular filters; the dialect result concerns classifier labels, not writers.
The EU summary is an inventory of what a model read.	The six commercial summaries examined name no individual website in the domain field, yet the same template carries a detailed list of named datasets, with links to reproduce the full training data, in the Apertus summary: the form is not what limits the detail.
robots.txt is a law, or a lock.	It is a request about future crawling. A 2024 audit of 14,000 web domains found that, within a single year, domains accounting for roughly 5% or more of C4's tokens had fully restricted at least one of the AI crawlers studied, restrictions that matter, in the authors' condition, "if respected or enforced".
The settlement proves infringement, or proves nothing.	It resolves the class's claims without deciding liability; the earlier summary-judgment order is a separate record.
The enforcement date caused the publication dates.	The dates are consistent with that reading, and no evidence establishes it.
A revised disclosure proves concealment.	One summary was revised without a version increment; both texts remain downloadable, and no reason for the change is on the public record.

Disclosure: the system that produced this text was built by one of the companies whose summary is examined above.

11. Sources

Source	Date	Identifier or location
Regulation (EU) 2024/1689, Articles 53, 101, 111(3), 113; Commission explanatory notice and template	OJ 12 Jul 2024; template 24 Jul 2025; read 1 and 11 Sep 2026	eur-lex; artificialintelligenceact.eu (Art. 111); digital-strategy.ec.europa.eu
California AB 2013 (Stats. 2024, ch. 817), Civil Code §3111; *X.AI LLC v. Bonta*, C.D. Cal. 2:25-cv-12295, Dkt. 35; 9th Cir. No. 26-1591	approved 28 Sep 2024; order 4 Mar 2026; read 11 Sep 2026	leginfo.legislature.ca.gov; CourtListener
OpenAI, *Training Data Summary Pursuant to California Civil Code Section 3111*	Internet Archive capture, 21 Jan 2026	help.openai.com
Blankvoort, Pandit, Gahntz, *Quality Assessment of Public Summary of Training Content …*, and their tracker	arXiv v1, 26 Feb 2026 (FAccT 2026); tracker read 1 and 11 Sep 2026, last modified 31 Aug 2026	arXiv 2603.13270 v1; aial.ie
OpenAI, *Public Summary of Training Content* for GPT-5.5 (both live copies), GPT-5.6 Luna and GPT-6 Astra	GPT-5.5 v1, 29 Jul 2026 and v1, 26 Jun 2026, both re-fetched 11 Sep 2026; Luna v1, 23 Jul 2026; GPT-6 Astra v1, 2 Sep 2026, fetched 11 Sep 2026	cdn.openai.com
Anthropic (Claude Opus 5), Google (Gemini 3 Pro family), xAI (Grok 4.5), Inkling, Meta (Muse Spark) training-content summaries	23 Jul, 2 Jul, 8 Jul, 15 Jul and 4 Aug 2026, each from the document's own date field	providers' sites; AIAL archive
Swiss AI Initiative, *Apertus EU Public Summary*	V1, 1 Sep 2025	huggingface.co/swiss-ai
Francis and Kučera, *Manual of Information … A Standard Corpus of Present-Day Edited American English*	1964; rev. 1971; rev. and amplified 1979 (text of the 1979 revision)	Internet Archive capture of the ICAME edition, 13 Mar 2024
Radford, Wu, Child, Luan, Amodei, Sutskever, *Language Models are Unsupervised Multitask Learners*	Feb 2019	cdn.openai.com; no arXiv identifier exists
Raffel, Shazeer, Roberts, Lee et al., *Exploring the Limits of Transfer Learning …*	arXiv v1, 23 Oct 2019; v3, 28 Jul 2020; v4, 19 Sep 2023	arXiv 1910.10683
Gao, Biderman, Black et al., *The Pile*	arXiv v1, 31 Dec 2020	arXiv 2101.00027 v1
Dodge, Sap, Marasović et al., *Documenting the English Colossal Clean Crawled Corpus*	arXiv v1, 18 Apr 2021	arXiv 2104.08758 v1
Lee, Ippolito, Nystrom, Zhang, Eck, Callison-Burch, Carlini, *Deduplicating Training Data …*	arXiv v1, 14 Jul 2021	arXiv 2107.06499 v1
Carlini, Ippolito, Jagielski, Lee, Tramèr, Zhang, *Quantifying Memorization …*	arXiv v1, 15 Feb 2022	arXiv 2202.07646 v1
OpenAI, *GPT-4 Technical Report*	arXiv v1, 15 Mar 2023	arXiv 2303.08774 v1

Source	Date	Identifier or location
Penedo, Kydlíček, Ben allal et al., *The FineWeb Datasets*	arXiv v1, 25 Jun 2024	arXiv 2406.17557 v1
The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text	arXiv v1, 5 Jun 2025	arXiv 2506.05209 v1
Longpre, Mahari, Lee, Lund et al., *Consent in Crisis*	arXiv v1, 20 Jul 2024	arXiv 2407.14933 v1
Wan, Klyman, Kapoor, Maslej, Longpre, Xiong, Liang et al., *The 2025 Foundation Model Transparency Index*	2025 report; finalisation and release Sep–Dec 2025	crfm.stanford.edu/fmti
Bartz v. Anthropic PBC, Dkt. 231 and Dkt. 680	23 Jun 2025; 20 Jul 2026	N.D. Cal. 3:24-cv-05417, via RECAP
Statement of Interest of the United States, *In re OpenAI, Inc. Copyright Infringement Litigation*, 25-md-3143 (S.D.N.Y.)	1 Sep 2026	CourtListener, MDL docket entry 1682
Common Crawl overview, errata, and top-500 domains of CC-MAIN-2026-34	read 1 Sep 2026	commoncrawl.org; commoncrawl.github.io
nytimes.com/robots.txt and reddit.com/robots.txt, live and archived	read 1 and 11 Sep 2026	Internet Archive