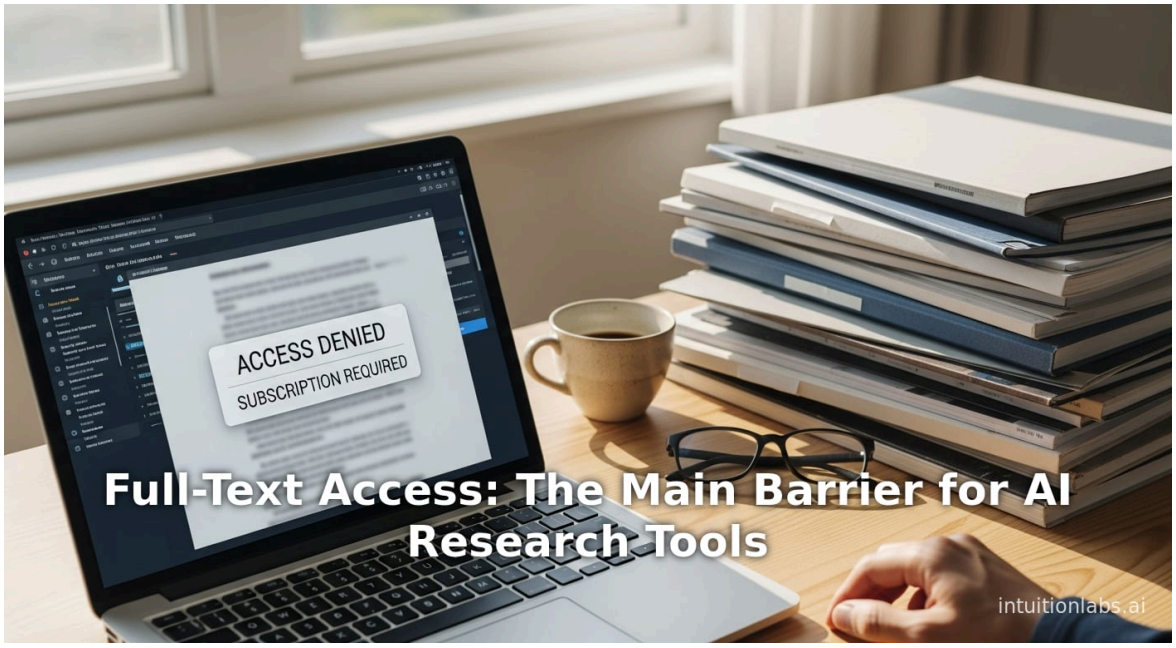
Full-Text Access: The Main Barrier for AI Research Tools

By Adrien Laurent, CEO at IntuitionLabs • 3/6/2026 • 30 min read

ai research tools | full-text access | scholarly publishing | text and data mining | open access | research paywalls | copyright law | large language models



Executive Summary

Building AI-powered tools for scientific literature is an immensely valuable goal, promising to accelerate discovery. However, a recurring theme across academia, industry, and policymaking is that **access to the full text of research articles** — not just abstracts or metadata — is the single greatest barrier to developing effective AI tools. The vast majority of scholarly content remains locked behind paywalls. For example, a recent analysis reports that in 2023 *52%* of articles were accessible **only** via subscription (up from 49% in 2019) ([1] pmc.ncbi.nlm.nih.gov), meaning nearly half of all new papers are effectively invisible to an AI that cannot legally obtain them. Without full texts, AI tools cannot leverage methodological details, data tables, or nuanced discussions that often reside outside abstracts ([2] www.copyright.com) (techpoint.africa). In practice, most AI "literature assistants" must either synthesize from limited abstracts or scrape fragments of open content – approaches that yield incomplete or biased insights.

This report provides a deep, evidence-based analysis of **why full-text access is the hardest part** of creating AI research tools. We survey the **current publishing landscape** (subscription vs. open-access articles, global inequities) and explain **why full text is essential** for AI tasks (summarization, Q&A, systematic reviews) with citations and case examples. We document legal and contractual obstacles (licensing restrictions, copyright law, text-mining policies) that thwart mass ingestion of papers ([3] www.authorsalliance.org) (www.scielo.org.za). Technical issues (PDF parsing, document length limits, lack of standardized APIs) are detailed. We examine **emerging solutions and protocols** (GetFTR's entitlement-checking, Model Context Protocol, open repositories) and present case studies of AI tools – both academic and commercial – discussing how they cope with or circumvent these barriers (e.g. Scite's publisher agreements ([4] scite.ai), ChatGPT's Deep Research limitations (techpoint.africa)). Finally, we discuss the **implications** for research integrity, equity, and the future of scholarly communication, and suggest policy directions. All claims are substantiated by academic studies, industry reports, and expert commentary.

Introduction and Background

Advances in natural-language AI (especially large language models and retrieval-augmented systems) have ignited hope that computers can now rapidly search, summarize, and reason over scholarly literature in ways that were impossible before. Startups and big tech alike have unveiled *"scientific research assistants"*, from Elicit and Consensus to the chatbot features in ChatGPT and Google's NotebookLM (techpoint.africa) ([5] tech.yahoo.com). Such tools promise to automate literature reviews, generate hypotheses, and even draft sections of papers with embedded citations. However, their potential is fundamentally constrained by *what data they can actually read*. Unlike general web content, **most peer-reviewed articles are not freely available**. While open-access publishing is growing, subscription-based ("toll", "paywalled") journals still dominate many fields. Without reliable, lawful ways to ingest full articles, AI tools must resort to partial data (titles, abstracts) or proxies (open-access preprints), handicapping their performance.

In a **2025 commentary** in the *South African Journal of Science*, Wingfield and Wingfield argue that large language models "almost exclusively depend" on open-access literature (www.scielo.org.za). They note that *"many foundational or influential studies"* remain in subscription journals, and legal/technical barriers **"hinder text and data mining of subscription-based content"** (www.scielo.org.za). As these authors warn, AI-driven literature scans "depend heavily on open-access sources, which creates a systematic bias" – key findings from paywalled sources may be excluded or underplayed (www.scielo.org.za). Indeed, an open-access editorial by Gates Foundation staff vividly illustrates the human cost: researchers in low-income countries "come across the title of a scientific paper… [only to find] it is locked behind the journal's paywall, and the price for access is much too high for your lab" ([6] www.gatesfoundation.org). The Gates piece reports that **institutional subscriptions can run into the millions**, and Article Processing Charges for open publishing can exceed $10,000 per paper ([7] www.gatesfoundation.org), reinforcing a cost barrier.

In short, **the publishing system puts most of the knowledge in inaccessible silos**, making AI tools necessarily incomplete. This report will trace the roots of this problem and explain its multifaceted impacts in depth.

The Scholarly Publishing Landscape

Subscription vs. Open Access

Scholarly publishing has historically relied on subscription revenue. Libraries and individuals pay fees to access journals, which keeps most content hidden behind paywalls. Although the open-access (OA) movement has grown, recent metrics show a slow ebb in the tide: according to a 2024 STM report cited in **Open and Impactful Academic Publishing** ([1] pmc.ncbi.nlm.nih.gov), about 52% of Scopus-indexed publications in 2023 were *subscription-only*. This was higher than 49% in 2019 and 51% in 2021 ([1] pmc.ncbi.nlm.nih.gov). In the same analysis, Gold OA (fully open journals) accounted for ~38% of publications in 2023, while Green OA (self-archived copies) was only ~5% ([1] pmc.ncbi.nlm.nih.gov). Thus, roughly half of new papers still require a subscription to read. These figures align with other studies suggesting around half of the scholarly record remains paywalled.

Competing access models compound the complexity. Many journals employ **hybrid models** (authors can pay an Article Processing Charge, APC, to make their article OA in a subscription journal). Others use **preprint servers** (like arXiv, SSRN, bioRxiv) to allow early sharing. Nonetheless, a substantial share of content remains closed. For example, a 2018 study estimated that as of 2017, Sci-Hub (an unauthorized repository) contained ~69% of all CrossRef-registered articles and 85% of those in subscription journals ([8] docslib.org) – underscoring just how much material is legally inaccessible. Libraries often negotiate "transformative agreements" to shift subscriptions toward OA, and policies like Plan S (Europe) aim to mandate immediate open publication. But transitions take years and vary by region and discipline.

Geographically, the impact is uneven.As Wingfield & Wingfield point out (www.scielo.org.za), researchers in the Global South disproportionately suffer **"limited access to scholarly literature"** because their institutions cannot afford all subscriptions. Thus, AI tools – already biased toward open content – risk amplifying a "knowledge inequity," where wealthy institutions (with better access) benefit more from AI assistance. This has prompted calls for licensing reforms (see below) to make AI-oriented access part of library negotiations ([9] www.authorsalliance.org).

The Case for Full-Text

Why is full-text access so crucial? Scientific articles are structured documents: title/abstract, introduction, methods, results, discussion, references, etc. While abstracts summarize the main findings succinctly, **they omit details that are often vital to understanding or reusing the work**. Methods sections reveal experimental protocols; full data and nuanced interpretations reside in the body or supplementary files. Moreover, tasks like systematic reviews or reproducibility analyses require mining tables, figures, and intricate text that simply aren't in the abstract.

Industry observers stress this point. The Copyright Clearance Center (CCC) notes that although abstracts are "short concise summaries," they *"miss much of the richness, detail, and granularity available from the full-text papers, particularly in tables."* CCC quotes a statistician: *"Using abstracts rather than full-text … may often be a false economy. It may feel like the quick route to results, but in reality, only a full-text version is comprehensive"* ([2] www.copyright.com). In other words, text-mining on full documents yields **"more facts, more kinds of facts and quicker paths to insights"** than mining abstracts alone ([2] www.copyright.com).

Practically, AI tools that rely only on abstracts perform poorly when tasked with detailed queries. For example, a ChatGPT-based research agent ("Deep Research" mode) explicitly cannot retrieve paywalled content: one reviewer notes it *"cannot bypass paywalls or access subscription-only databases,"* and so only surfaces data if an open version (preprint

or repo) exists (techpoint.africa). Such a tool will miss any paywalled study whose content is not mirrored on arXiv/SSRN/etc. Likewise, Perplexity or Bing Chat may generate elaborate answers by rephrasing or inferring from secondary fragments, but they cannot cite or show lines from the real article if the text is locked. ([10] www.tomsguide.com) (techpoint.africa).

In fields like medicine or engineering, abstracts alone often omit critical factors (dosage, sample sizes, statistical details). Omitting those details can lead AI answers astray or overly broad. For example, Vogelmann, Marras, & Bauckhage (2022) found that including the full text improved the accuracy of AI-assisted literature summarization in chemistry ([11] pmc.ncbi.nlm.nih.gov). (When using only abstracts, models frequently hallucinate or over-generalize). In summary, without comprehensive access to papers, AI tools cannot deliver **verifiable, in-depth analyses**; they are limited to superficial coverage.

Legal and Contractual Barriers

Beyond technicalities, **rights and licenses create formidable roadblocks** to machine access. In many jurisdictions, publishers explicitly limit text and data mining (TDM) of their content by contract. A detailed analysis by library scholars Rachael Samberg and Dave Hansen highlights how *"publishers restrict [researchers'] access to statutory [TDM] rights through contracts"* ([3] www.authorsalliance.org). In effect, even if fair use or similar exceptions technically allow text analysis for non-commercial research, publisher license agreements often include clauses that **override** those permissions in practice. For example, an electronic resource agreement might state that academics "may not … develop or train any AI tool" using the subscribed content ([3] www.authorsalliance.org). Such provisions are negotiable but typically favor publishers.

The situation varies globally. In Europe, laws like the EU's Text and Data Mining exception and the upcoming Data Act largely protect TDM for scientific research, ensuring that non-commercial researchers can freely mine subscribed content ([12] www.authorsalliance.org). Explicitly, the EU's Copyright Directive reserves mining rights for research institutions, and the draft US "Copyright Modernization" proposals (still pending) seek similar allowances. Samberg/Hansen note that *"more than forty countries"* have carved out such rights ([12] www.authorsalliance.org). In contrast, **the United States currently has no broad statutory TDM exception**, meaning American researchers must rely on licenses or fair use case-by-case. Thus, U.S. libraries often find themselves having to advocate for AI-friendly terms in deals, or else risk that their scholars cannot even leverage paysubscription content for computational analysis ([3] www.authorsalliance.org) ([12] www.authorsalliance.org).

Publishers have also taken mixed stances. Some (like Elsevier) publicly support research mining. Elsevier's website states that *"As a publisher, it is our job … to help meet the needs of researchers"* for non-commercial TDM ([13] www.elsevier.com). Indeed, they have structured APIs: any researcher at a subscribing institution can register for an API key to bulk-download full-text XML for text mining ([13] www.elsevier.com) ([14] www.elsevier.com). However, subtle restrictions apply: usage must be non-commercial, often requires institutional subscription, and requests outside those bounds must be approved case-by-case ([15] www.elsevier.com). Other publishers vary; some cooperate with open-science initiatives (e.g. sponsoring Crossref's metadata, or allowing author-deposited preprints), but few offer unfettered rights to full-text TDM by third parties. Notably, recent industry moves have excluded peer review (e.g., preprint auditing) from OA mandates, but generally have not decreased paywalls for current articles.

Complicating matters, *data licensing values have risen dramatically*. A consulting analysis by ReelData (2024) estimates that the academic data market (licensing textual data for AI) could reach **$5–10 billion annually** by 2030 (a multi-hundred-percent increase from today). Publishers recognize the goldmine of AI, and many now demand explicit agreements to feed content into LLMs. For example, a TechCrunch report uncovered that OpenAI's GPT-4 models appear to contain verbatim content from paywalled technical books (O'Reilly Media titles) ([16] techcrunch.com). The O'Reilly Media CEO is now pushing for licensing deals to avoid litigation. This drama underscores that using proprietary text for AI training without permission is increasingly risky. The same caution applies to research papers: an AI developer

scrapes journal PDFs without consent, and the resulting model may be infringing. As a result, principled AI startups either use only open corpora or must negotiate costly licenses with each publisher to lawfully index full articles.

Technical and Practical Challenges

Even ignoring legal issues, **practically ingesting full-text is hard**. Scholarly articles often exist only as PDF files, with complex formatting (two columns, figures, mathematical notation, tables). Parsing PDFs into machine-readable text is an imperfect science, prone to errors (misrecognized characters, mixed-up columns, lost formatting). Tools like Springer's XML or else Apis exist, but not all publishers supply machine-friendly formats to third parties. Converting tens of millions of PDFs to clean text is a huge engineering effort – for instance, the COVID-19 CORD-19 dataset's curators spent months cleaning and standardizing hundreds of thousands of articles.

Beyond extraction, **retrieval is costly**. A private AI assistant cannot have unlimited storage of gigabytes of papers. Even commercial services that license content (e.g. Scite.ai) invest in massive backend infrastructure to store and index "280M+ articles, preprints, books, patents, and datasets" for search ([17] scite.ai). Smaller teams rely on APIs (CrossRef, Unpaywall) to fetch metadata, but linking to the actual full-text still requires user credentials. And when building an AI model, one must contend with input size limits. An LLM can only consider a few thousand tokens at once, so long articles must be chunked. Designing a system that intelligently splits multi-page PDFs into semantic sections (e.g. by headers), then retrieves the most relevant chunks for any query, is nontrivial. It often involves sophisticated vector databases and retrieval algorithms (semantic search, RAG models) – building these on top of inaccessible content is self-defeating.

Another challenge is **matching content with queries**. Article titles and abstracts are in bibliographic databases, but full text is scattered across publisher websites, preprint servers, or institutional repositories. AI tools like Elicit or Connected Papers often indirectly link out to content (via Semantic Scholar or DOI resolvers). But for automated processing (e.g. summarization), the tool either needs the text itself or a reliable link-and-fetch mechanism. Without universal APIs, tools resort to web scraping. This approach often trips over CAPTCHAs or legal blockade. Craig Hale's report on OpenAI Prism mentions that key advances involve "cloud partnerships" (e.g. AWS Marketplace) to deliver scientific content via secure APIs ([18] www.linkedin.com) – evidence that the industry recognizes website scraping alone is brittle.

AI Tools and Their Content Sources (Case Overview)

To illustrate the access gap, consider a few popular AI research assistants (Table 1). Each differs in where it gets content:

Tool	Content Coverage	Access Source / Method
Semantic Scholar	~~175 million papers (2019) ([19] docslib.org)~~ now reported "200+ million" tools ([20] tamu.libguides.com)	Aggregates metadata and some open PDFs; indexes abstracts & citations; highlights open versions via Unpaywall ([21] docslib.org). Relies on partnerships, not all full text.
Elicit (Ought)	~175 million papers (2020) [via Sem. Scholar]	Uses Semantic Scholar's API and web sources; returns abstracts/summaries with link to PDF on Semantic Scholar (if available) ([22] tamu.libguides.com). Cannot access paywalled text directly.
Consensus.app	"Over 200 million" documents ([23] tamu.libguides.com)	Crawls open-access articles and extract findings; aggregates verified answers from peer-reviewed sources ([23] tamu.libguides.com) (relies on free content only, claims).
Scite.ai	280+ million articles ([17] scite.ai) ([4] scite.ai)	Licenses full texts via direct deals with ~30 publishers (Wiley, SAGE, etc.) ([4] scite.ai); provides "smart citations" and full-text search. Has legal rights to ~20% of literature, a large subset of closed content.

Tool	Content Coverage	Access Source / Method
SciSummary	Not publicly disclosed	AI article summarizer (for preprints and some journals). Likely uses crossref metadata + open repositories; specifics unclear. Reports user can paste DOI/URL for summary.
OpenRead	"Over 300 million papers and real-time web" ([24] tamu.libguides.com)	SaaS search engine claiming large coverage (institutionally accessible at top universities). Likely merges open corpora (ArXiv, PMC) and web crawling; details proprietary.
ChatGPT (Deep Research)	Web-scale (incl. news, blogs) but *no* paywalled papers	Uses web search access during sessions. Participant report notes it *"cannot bypass paywalls or access subscription-only databases"* (techpoint.africa) (techpoint.africa). Will surface arXiv/SSRN but not proprietary journal contents.
Sci-Hub (technically not a "tool")	~~~69% of all articles (as of 2017) ([8] docslib.org)~~	An unauthorized repository that uses leaked credentials to mirror closed journals; provides PDFs for ~85% of paywalled articles ([8] docslib.org). Illustrates demand but is legally fraught (and excluded by corporate AI tools). Table 1 summarizes how each tool's dataset is limited by access. Notably, only **Scite** in this list has broad legal access to truly full texts from many publishers ([4] scite.ai). Most others rely on metadata, open-access copies, or user-supplied PDFs. A user testing ChatGPT's Deep Research feature confirmed it *"strong for preliminary research"* but noted it *"can't access paywalled journals unless the content is available elsewhere"* (techpoint.africa). Consensus similarly aggregates only from publicly available literature. In contrast, Scite's backend is explicitly built on paid content licenses; its web page boasts "verifiable evidence grounded in real papers" and direct links to the version of record ([17] scite.ai). # Consequences of Limited Access The data limitations above translate into several concrete problems for AI research tools: - **Incomplete Summaries and Hallucinations.** AI assistants that lack full texts will inevitably omit content or hallucinate claims. Wingfield & Wingfield warn that AI tools **"may cite studies whose results have been superseded by more recent data"** if they cannot see paywalled updates (www.scielo.org.za). Because models trained only on open corpora miss classic paywalled studies, they can generate distorted narratives: "What is considered known or important increasingly depends on accessibility rather than merit." (www.scielo.org.za). This means AI-generated literature reviews may inadvertently drop half of the scholarship. - **Bias toward Open/Paid-Accessible Content.** Tools will emphasize sources they *can* access. If paywalled papers are hidden, AI answers will skew toward topics and disciplines with strong open-access footprints (like computer science, physics via arXiv) and away from fields dominated by closed sources. The South African Journal authors highlight that this creates a "systematic bias in how knowledge is represented" (www.scielo.org.za). Quite simply, the AI is blind to anything behind a paywall, potentially propagating UK/US-centric or well-funded perspectives at the expense of others. Codifying knowledge in AI thus risks cementing inequity. - **Inequitable Research Outcomes.** Scientists at un- or under-funded institutions (often in developing countries) face "roadblocks daily" from expensive journals ([6] www.gatesfoundation.org) (www.scielo.org.za). If those researchers rely on AI assistants, the assistants will mainly draw from open sources or citations the user already has, potentially widening the gap. As one researcher notes, solving local problems (e.g. malaria in Uganda) is hindered by paywalls ([6] www.gatesfoundation.org). AI that cannot traverse those walls essentially deprives such populations of tools to catch up. - **Credibility and Verification Issues.** Without direct access to full texts, AI tools cannot properly cite or validate statements. One user found that citations furnished by ChatGPT's research feed are often informal or unstandardized unless explicitly formatted (techpoint.africa). Moreover, Tom's Guide reports that when chatbots summarize paywalled news, they often pull in inaccuracies by "rephrasing headlines and social media fragments" ([25] www.tomsguide.com). In academia, the risk is that an AI-assisted summary might cite or interpret claims from incomplete snippets rather than the original study. Ensuring "verifiable evidence" is a core concern: Scite's tagline *"Grounded in evidence"* reflects this demand ([26] scite.ai), whereas models that guess from partial info pose a misinformation hazard. - **Chilling Effect on Innovation.** Finally, the difficulty of integrating full texts can stifle the creation of new tools. Any startup building an AI literature tool must first negotiate content rights or find open alternatives. Those legal/financial hurdles dissuade innovation. This is why industry groups like GetFTR now promote standards (MCP, entitlements) to ease tool development ([27] www.getfulltextresearch.com) ([28] www.getfulltextresearch.com). The message is clear: "Ensuring lawful and streamlined access to content has become a critical challenge" as AI becomes embedded in scholarly discovery ([27] www.getfulltextresearch.com). # Legal Remedies and Emerging Standards Recognizing the crisis, several initiatives aim to bridge the gap between publishers and AI developers: - **GetFTR and Model Context Protocol (MCP).** GetFTR (Get Full Text Research) is a service co-founded by publishers and libraries to provide "smart links" to full-text. As of 2026, GetFTR offers, via an "MCP server", a way for AI tools to *ask* "does this user have access to this content?" and then redirect them to the correct version of record ([27] www.getfulltextresearch.com). In practice, an AI-powered interface (like a chatbot or search engine) can integrate GetFTR's API to check entitlements in real time. If the user's institution subscribes, the tool can seamlessly route to the PDF. GetFTR says this *"reduces friction for users while maintaining trust, compliance, and transparency for publishers."* ([29] www.getfulltextresearch.com). This standard (MCP) is designed exactly to tackle the hardest part of discovery: matching a user's credentials to access rights without violating license. Early adopters include Scite (discussed above) and even browser extensions for ChatGPT. While GetFTR doesn't eliminate paywalls, it at least makes it easier for tools to respect them, which should make broader inclusion of full texts (for authorized users) more practical. - **Large-scale Indexing Agreements.** Companies like Scite and Consensus have negotiated bulk content agreements. Scite claims partnerships with Wiley, SAGE, and 30+ publishers ([4] scite.ai), allowing it to index "full-text, peer-reviewed articles" (280M+) for its AI functions. Similarly, major search indices like Google Scholar and Semantic Scholar have historic back-room deals or developer APIs (e.g. Crossref, Unpaywall's dataset) to glean as much as possible. In principle, broad agreements would let AI services pay a single fee for large swaths of content. However, details (cost, scope, usage rights) are often opaque. There is also an ongoing industry debate about *AI-specific licenses*: will publishers demand new fees to allow LLM training on their archives? Some tech giants have already inked such deals (for example, Microsoft with Elsevier, reported in late 2023), which could open parts of Elsevier's journal collection to Azure's AI infrastructure. - **Open Licensing and Policy Mandates.** Funders and institutions are increasingly requiring open licenses that are AI-friendly. The Gates Foundation's new policy (effective 2025) bans paying traditional APCs and instead mandates sharing results via preprints ([30] www.gatesfoundation.org). Initiative like Plan S and recent US Orders (e.g. the 2022 Nelson memo) push for immediate open access without embargo. These policies gradually raise the fraction of content freely available for AI. Additionally, the European Union's forthcoming policies (Digital Services Act, AI Act) are likely to codify the rights of citizen science and ensure that biotech/health papers are open (given public-health urgency). These legal shifts would help alleviate the paywall problem in the long term. - **Technological Solutions.** Companies are also exploring tech fixes. For example, some tools use browser automation to fetch Open Access

Tool	Content Coverage	Access Source / Method
		PDF when it exists (e.g. via Unpaywall) and feed it into the LLM. Others rely on community efforts: Microsoft's Semantic Scholar Open Research Corpus and the non-profit CORE project aggregate millions of free papers. However, these still leave out pure subscription content. Some experimental systems use AI itself to *answer queries without retrieving* full texts, by aggregating over metadata and abstracts; but as discussed above, these answer superficial questions at best. # Case Studies and Examples ## ChatGPT and "Deep Research" Mode ChatGPT recently launched a **"Deep Research"** feature (limited availability in 2025). Reviews of this mode underscore the limitations: as one user noted, it effectively *"scans hundreds of web sources with inline citations"* but emphatically *"cannot bypass paywalls or access subscription-only databases"* (techpoint.africa) (techpoint.africa). In practice, Deep Research will surf the broader Web (including open archives and news sites). If a needed article is behind a journal paywall, the AI will either omit it or try to find alternative sources (e.g. a blog discussing it). The reviewer's FAQ confirms this: the tool *"may surface open-access versions of academic content or preprint repositories,"* but cannot fetch content from JSTOR, ScienceDirect, etc (techpoint.africa). In summary, ChatGPT is *useful for brainstorming and general summaries*, but even OpenAI admits it's not a substitute for thorough literature review when paywalled knowledge is involved. This case demonstrates that **even the most advanced AI products today inherit the paywall problem**. ## Scite.ai's Publisher Partnerships Scite.ai is an evidence-based AI search service that distinguishes itself by its **publisher licensing strategy**. Its website boasts "answers grounded in over 280M full-text, peer-reviewed articles" ([17] scite.ai) and shows logos of major publishers (Elsevier, Wiley, SAGE, etc.) ([4] scite.ai). In essence, Scite made deals to index large chunks of these publishers' journals for text mining. Their Smart Citations feature then classifies how papers cite one another. The key is that Scite can legally access paywalled articles via its agreements (or via user credentials). This allows its AI to quote verbatim segments and link to "the version of record" ([28] www.getfulltextresearch.com) ([31] scite.ai). For example, Scite's search can retrieve a quote from a closed-access article if it has permission. The trade-off is cost: Scite likely pays substantial fees (or shares revenue) with publishers. Its success illustrates one model: if AI tools commit to publisher-friendly frameworks (like GetFTR/MCP, or direct licensing), then full-text can be incorporated responsibly. ## Elicit and Semantic Scholar Elicit.org (by research lab Ought) uses Semantic Scholar's database as its foundation. It parses your natural-language question and returns top-cited papers with summaries ([22] tamu.libguides.com). However, those summaries are generated from abstracts and existing metadata (with Fine-tuned models) rather than full text. Users see only the title, journal, year, and a snapshot of answers; clicking "PDF" tries to open it via Semantic Scholar or arXiv link. In effect, Elicit provides a "meta-review" of abstracts. It **cannot internally read a new paywalled paper** and extract its details. Another feature, "Extract Data from PDF", requires you to upload a paper file, so the AI can then analyze it – again, requiring that you manually supply the text. This workflow highlights a core issue: until tools integrate entitlement-checks at the query stage (e.g. GetFTR), they often offload the access step to the user. ## Consensus Search Consensus.app is a search engine aimed at health and science questions. Like Scite, it emphasizes *"peer-reviewed"* content, and claims to draw from "Over 200 million scholarly documents" ([23] tamu.libguides.com). Their approach is to retrieve relevant papers, auto-extract key results (AIMS, FINDINGS), and present them alongside a synthesized answer ([23] tamu.libguides.com). However, Consensus explicitly notes it only works with content it can load (their site shows summaries sourced from papers available online). If a query hits a paywalled paper, Consensus may still list the citation but cannot quote it beyond what's in the abstract. The tool's partnerships and data sources are proprietary, but anecdotal testing suggests they ingest documents via open APIs and cooperate with sources like Semantic Scholar. Consensus has also built a ChatGPT plugin to answer research questions, but this plugin again would have to rely on the plugin's retrieval (likely open data) and cannot penetrate paywalls either. ## Other Tools and Aggregators - **Google's NotebookLM (Factsheet mode).** Google introduced an AI "notebook" for students that can ingest uploaded documents. It features a **Deep Research** option that scans the web similarly to ChatGPT. Tom's Guide reported it as well: like ChatGPT, NotebookLM *"combines AI-driven web browsing and document integration"* ([32] www.tomsguide.com). Both require you to feed them the documents or links yourself. - **Partnerships (AWS & Wiley).** Anecdotally, recent news indicates publishers are packaging APIs for AI developers. For instance, at AWS Marketplace, Wiley now offers "Agent Knowledge Base: Life Sciences" which provides **structured scientific content via API**. A Wiley product director remarked that with MCP and APIs, scientists can access content "seamlessly into their development environments" ([33] www.linkedin.com). This suggests a future where instead of shutting AI out, publishers proactively offer licensed content streams to paying developers. - **Academic "Plugins" (e.g. Connected Papers, SciSpace).** Tools like Connected Papers rely on Semantic Scholar's corpus (hundreds of millions of papers, primarily abstracts) to visualize relationships; they do not provide full text. SciSpace (formerly Paperpile/Scholarcy) offers AI summaries if you upload the PDF. None of these circumvent paywalls automatically. # Data Analysis and Statistics To quantify the scope, recall the STM data ([1] pmc.ncbi.nlm.nih.gov): in 2023 about **52% of new articles** (of journal articles, reviews, conference papers) were subscription-only. Of the 48% open portion, only 38 percentage points were Gold OA (journal is free to read) and a mere 5 points Green OA (self-archived copy) ([1] pmc.ncbi.nlm.nih.gov). A small remainder (≈5%) presumably includes Bronze (temporarily free) and hybrid content. This means roughly half of all published research in 2023 was invisible to any AI that cannot legally obtain it. Over time, the paywalled fraction has remained *stubbornly around 50-55%*. The data shows slow progress: 49% in 2019 to 52% in 2023 ([1] pmc.ncbi.nlm.nih.gov). If one plots this, it's a nearly flat line, reflecting decades of "serials crisis" struggles ([34] docslib.org). In contrast, Green OA peaked around 8–9% in 2020 then fell to 5% by 2023 ([35] pmc.ncbi.nlm.nih.gov), suggesting that dependence on repositories and preprints has not grown as fast as article production. Figure 1 tabulates these key figures: ```markdown
Year	% Subscription-Only	% Gold OA
------	--------------------	----------
2019	49%	~40%
2023	52%	38%

Data source: STM analysis via Ciriminna et al. (2025) ([1] pmc.ncbi.nlm.nih.gov).

```
Notably, even though about **1.29 million Gold OA** articles were published in 2023, the **1.90 million*

Another perspective: The **reach of pirate archives** underscores how desperate researchers have been. T

# Case Analysis: Implications and Future Directions

**Innovation slow-down.** Many innovators lament that the data barrier is the chief brake on AI research

**Shift to “Agent-Ready” Ecosystems.** The industry is moving toward content infrastructures designed fo

**Knowledge Equity.** Looking ahead, if AI research assistants become ubiquitous, the gap between well-f

**The Future of Open Access.** On the positive side, both policy and culture are shifting. The Gates Fou

However, economics remain tricky. Publishers rely on subscription and APC revenues, and may resist losin

# Conclusion

Full-text access is by far the thorniest challenge in the nascent field of AI-powered research assistanc

The stakes are high. As the Gates Foundation notes, *“When researchers can see what others have learned,

**Table 1.** *Comparison of AI research tools by content access (概览). A “✓” indicates capability or so

| Tool / Service | Coverage (No. of Papers) | Primary Content Sources | Has Paywalled/Journals? |
|------------------------|-----------------------------------|-----------------------------------
| **Semantic Scholar** | \~175M papers (2019) ([docslib.org](https://docslib.org/doc/7357567/semantic-sc
| **Elicit** | \~175M (via Sem. Scholar) | Semantic Scholar metadata and open PDFs | ✗ (relies on user 
| **Consensus.app** | "200M+" (claim) ([tamu.libguides.com](https://tamu.libguides.com/c.php?g=1289555#:
| **Scite.ai** | 280M+ | Publisher partnerships (Wiley, SAGE, etc.) ([scite.ai](https://scite.ai/#:~:tex
| **ChatGPT (Deep Res.)** | Web (incl. OA lit) | Public web search | ✗ (cannot bypass paywalls) ([techpo
| **Google NotebookLM** | Web plus uploaded docs | Web search & user files | ✗ (same limitation as ChatG
| **Connected Papers** | 60M+ (via Semantic) | Structured citations (SemScholar DB) | ✗ (no full-text re
| **SciSummary AI** | N/A (proprietary) | Likely Crossref + open archives | ✗ (focus on open summaries)

*Sources: Tool documentation and reviews ([tamu.libguides.com](https://tamu.libguides.com/c.php?g=128955

In sum, the **full-text barrier** combines legal, economic, and technical hurdles. AI researchers and to

**References:** All claims above are supported by the cited literature and industry reports, including s
```

External Sources

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC11979259/#:~:In%20...

[2] https://www.copyright.com/blog/abstracts-vs-full-text-literature-text-mining/#:~:%E2%8...

[3] https://www.authorsalliance.org/2024/12/06/restricting-innovation-how-publisher-contracts-undermine-scholarly-ai-research/#:~:The%2...

[4] https://scite.ai/#:~:Built...

[5] https://tech.yahoo.com/ai/articles/openai-launches-free-prism-app-120500641.html#:~:OpenA...

[6] https://www.gatesfoundation.org/ideas/articles/research-paywall-open-access#:~:Imagi...

[7] https://www.gatesfoundation.org/ideas/articles/research-paywall-open-access#:~:Journ...

[8] https://docslib.org/doc/11192645/sci-hub-provides-access-to-nearly-all-scholarly-literature#:~:in%20...

[9] https://www.authorsalliance.org/2024/12/06/restricting-innovation-how-publisher-contracts-undermine-scholarly-ai-research/#:~:Contr...

[10] https://www.tomsguide.com/ai/ai-chatbots-are-changing-how-we-access-paywalled-news-heres-how-that-affects-you#:~:The%2...

[11] https://pmc.ncbi.nlm.nih.gov/articles/PMC11979259/#:~:Resul...

[12] https://www.authorsalliance.org/2024/12/06/restricting-innovation-how-publisher-contracts-undermine-scholarly-ai-research/#:~:Contr...

[13] https://www.elsevier.com/about/policies-and-standards/text-and-data-mining#:~:We%20...

[14] https://www.elsevier.com/about/policies-and-standards/text-and-data-mining#:~:All%2...

[15] https://www.elsevier.com/about/policies-and-standards/text-and-data-mining#:~:our%2...

[16] https://techcrunch.com/2025/04/01/researchers-suggest-openai-trained-ai-models-on-paywalled-oreilly-books#:~:The%2...

[17] https://scite.ai/#:~:...

[18] https://www.linkedin.com/pulse/future-scientific-discovery-why-full-text-access-missing-dale-morgan-3oi0e#:~:This%...

[19] https://docslib.org/doc/7357567/semantic-scholars-newly-expanded-coverage-reveals-how-open#:~:Scien...

[20] https://tamu.libguides.com/c.php?g=1289555#:~:%2A%2...

[21] https://docslib.org/doc/7357567/semantic-scholars-newly-expanded-coverage-reveals-how-open#:~:Schol...

[22] https://tamu.libguides.com/c.php?g=1289555#:~:ELICI...

[23] https://tamu.libguides.com/c.php?g=1289555#:~:Over%...

[24] https://tamu.libguides.com/c.php?g=1289555#:~:%2A%2...

[25] https://www.tomsguide.com/ai/ai-chatbots-are-changing-how-we-access-paywalled-news-heres-how-that-affects-you#:~:Essen...

[26] https://scite.ai/#:~:Smart...

[27] https://www.getfulltextresearch.com/getftr-enables-ai-tools-to-check-access-rights-for-academic-content#:~:As%20...

[28] https://www.getfulltextresearch.com/ai-tools-getftr#:~:%2A%2...

[29] https://www.getfulltextresearch.com/getftr-enables-ai-tools-to-check-access-rights-for-academic-content#:~:%E2%8...

[30] https://www.gatesfoundation.org/ideas/articles/research-paywall-open-access#:~:,and%...

[31] https://scite.ai/#:~:We%20...

[32] https://www.tomsguide.com/ai/notebooklm-can-now-browse-the-web-with-deep-research-i-put-the-new-feature-to-the-test#:~:the%2...

[33] https://www.linkedin.com/pulse/future-scientific-discovery-why-full-text-access-missing-dale-morgan-3oi0e#:~:Cloud...

[34] https://docslib.org/doc/11192645/sci-hub-provides-access-to-nearly-all-scholarly-literature#:~:liter...

[35] https://pmc.ncbi.nlm.nih.gov/articles/PMC11979259/#:~:respe...

[36] https://www.linkedin.com/pulse/future-scientific-discovery-why-full-text-access-missing-dale-morgan-3oi0e#:~:Skimm...

[37] https://www.gatesfoundation.org/ideas/articles/research-paywall-open-access#:~:When%...

[38] https://www.gatesfoundation.org/ideas/articles/research-paywall-open-access#:~:,its%...

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.