IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

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

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-03-06 · ai research tools · full-text access · scholarly publishing · text and data mining · open access · research paywalls · copyright law · large language models

Original article: https://intuitionlabs.ai/articles/full-text-access-barriers-ai-research-tools



In this article

1. Executive Summary
2. Introduction and Background
3. The Scholarly Publishing Landscape
4. Subscription vs. Open Access
5. The Case for Full-Text
6. Legal and Contractual Barriers
7. Technical and Practical Challenges
8. AI Tools and Their Content Sources (Case Overview)
9. Consequences of Limited Access
10. Legal Remedies and Emerging Standards
11. Case Studies and Examples
12. ChatGPT and “Deep Research” Mode
13. Scite.ai’s Publisher Partnerships
14. Elicit and Semantic Scholar
15. Consensus Search
16. Other Tools and Aggregators
17. Data Analysis and Statistics
18. Case Analysis: Implications and Future Directions
19. Conclusion

Executive Summary

Building AI-powered tools for scientific literature is an immensely valuable goal, promising to accelerate discovery. This report argues that **access to the full text of research articles** — not just abstracts or metadata — deserves priority when developing effective AI tools. Roughly half (52%) of Scopus-indexed journal articles, reviews and conference papers published in 2023 were subscription-only, compared with 49% in 2019 (Ciriminna et al., 2025). Without full texts, AI tools cannot leverage methodological details, data tables, or nuanced discussions that often reside outside abstracts (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 develops the argument that **full-text access deserves priority** when 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 (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 (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) (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, **access barriers leave gaps in the scholarly literature available to AI systems** (Wingfield and Wingfield, 2025). 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" (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 (www.gatesfoundation.org), reinforcing a cost barrier.

In short, **legal and technical barriers leave gaps in AI access to the scholarly record** (Wingfield and Wingfield, 2025). 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 subscription journals. 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** (pmc.ncbi.nlm.nih.gov), roughly half (52%) of Scopus-indexed journal articles, reviews and conference papers published in 2023 were *subscription-only*. This was higher than 49% in 2019 and 51% in 2021

(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% (pmc.ncbi.nlm.nih.gov). These percentages describe that Scopus-indexed publication set.

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. 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 (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"* (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 (www.copyright.com).

Practically, AI tools that rely only on abstracts perform poorly when tasked with detailed queries. For example, ChatGPT Deep Research's sources depend on its configuration. OpenAI's instructions allow relevant files and connected app context, and Consensus documents that it works inside ChatGPT Deep Research (official OpenAI documentation; Consensus changelog). Public-web search alone should not be treated as access to every subscription article; connected scientific-research services can route users according to their entitlements (Scite MCP documentation). 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. (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. Self-archiving is one route to broader access: Ciriminna and colleagues (2025) discuss its benefits for researchers and society (study). 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"* (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 (www.authorsalliance.org). Such provisions are negotiable but typically favor publishers.

In the EU, Article 3 of Directive (EU) 2019/790 covers text and data mining for scientific research by qualifying research organisations and cultural heritage institutions with lawful access. Copies must be stored with an appropriate level of security (EUR-Lex). Other users and uses require a separate assessment of applicable law and permissions.

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 (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 (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 (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.

Full-text access can involve publisher partnerships: Consensus documents access through publisher partnerships and open-access PDFs (Consensus documentation).

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 (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. Context limits should be checked for the selected model: for example, GPT-4.1 has a context window of about one million tokens (OpenAI documentation). Chunking should reflect that limit and the retrieval design. 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 (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) (docslib.org)~~ now reported "200+ million" tools (tamu.libguides.com)	Aggregates metadata and some open PDFs; indexes abstracts & citations; highlights open versions via Unpaywall (docslib.org). Relies on partnerships, not all full text.
Elicit (Ought)	Systematic Review workflows (Pro or higher for browser-extension retrieval)	Full-text screening and extraction in Systematic Review workflows on Pro or higher plans; the browser extension uses active institutional logins to retrieve papers the institution provides (Elicit documentation).
Consensus.app	Research-paper database; full-text coverage varies by paper	Analyzes participating publishers' full-text articles within the platform, alongside open-access PDFs. Viewing or downloading a publisher article separately depends on the user's institutional or personal subscription (Consensus documentation).

Tool	Content Coverage	Access Source / Method
Scite.ai	Licensed scholarly content and open-access corpus	Publisher agreements provide licensed scholarly content alongside the open-access corpus (Scite).
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" (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)	Public web, relevant files and connected app context, depending on configuration	OpenAI describes files and connected app context; Consensus works inside ChatGPT Deep Research (official OpenAI documentation; Consensus changelog). Scientific full-text routing depends on entitlements (Scite MCP documentation).

Table 1 compares content coverage and access methods. Access should be assessed for the particular tool and paper: Consensus notes that full-text access is not available for every paper (Consensus documentation).

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). Restricted access can leave gaps in the scholarly record available to AI tools (Wingfield and Wingfield, 2025).
- **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 (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 (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" (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 (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 (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 (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 (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."* (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's MCP searches inside article full text rather than only metadata (Research Solutions).
- **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 (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. The AI Act's rules on general-purpose AI became effective in August 2025 (European Commission).

- **Technological Solutions.** Companies are also exploring tech fixes. For example, some tools use browser automation to fetch Open Access PDF when it exists (e.g. via Unpaywall) and feed it into the LLM. Others rely on community efforts: the Semantic Scholar Open Research Corpus (S2ORC), developed at the Allen Institute for AI (official repository), 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

The configuration matters when assessing ChatGPT Deep Research’s access. Consensus announced on May 13, 2026 that its service works inside ChatGPT Deep Research (Consensus changelog). A connected research source should be assessed separately from the ability to bypass a publisher paywall.

Scite.ai’s Publisher Partnerships

Scite.ai is an evidence-based AI search service that distinguishes itself by its **publisher licensing strategy**. Its website describes full-text research and citation context alongside journal, organization, funder and patent data (Scite Data) and shows logos of major publishers (Elsevier, Wiley, SAGE, etc.) (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” (www.getfulltextresearch.com) (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 supports full-text screening and extraction in its Systematic Review workflows. Its browser extension can use active institutional browser logins to retrieve papers the institution makes available; users can upload remaining PDFs themselves (Elicit documentation). Automatic retrieval of subscription content depends on appropriate access, rather than providing unrestricted access to every subscription article.

Consensus Search

Consensus analyzes full text, including paywalled papers from major publishers (Consensus). This describes analysis inside the service; it should not be read as a promise of a downloadable PDF for every result.

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"* (www.tomsguide.com). Users can supply source material, but can also start with a research question: Google documents finding web sources from a question (Google documentation), while ChatGPT Deep Research accepts a question, scope and desired result, with relevant files or connected context (official OpenAI documentation).
- **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" (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 Typeset; rebranding announcement) 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 (pmc.ncbi.nlm.nih.gov): in 2023 about **52% of Scopus-indexed journal articles, reviews and conference papers published that year** 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) (pmc.ncbi.nlm.nih.gov). The 52% subscription-only figure describes roughly half of that indexed publication set, rather than all published research.

For Scopus-indexed journal articles, reviews and conference papers, the reported subscription-only share rose from 49% in 2019 to 51% in 2021 and 52% in 2023 (pmc.ncbi.nlm.nih.gov). If one plots this, it's a nearly flat line, reflecting decades of "serials crisis" struggles (docslib.org). In contrast, Green OA peaked around 8–9% in 2020 then fell to 5% by 2023 (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:

Year	% Subscription-Only	% Gold OA	% Green OA	% Other/Hyb.
2019	49%	~40%	6%	~5%
2023	52%	38%	5%	~5%

Population: Scopus-indexed journal articles, reviews and conference papers. Data source: STM analysis via Ciriminna et al. (2025) (pmc.ncbi.nlm.nih.gov).

Notably, even though about **1.29 million Gold OA** articles were published in 2023, the **1.90 million** that remained paywalled still outnumber them (pmc.ncbi.nlm.nih.gov). The subscription-only share was a small majority of that Scopus-indexed publication set in 2023.

Case Analysis: Implications and Future Directions

Innovation slow-down. Many innovators lament that the data barrier is the chief brake on AI research. As Dale Morgan of Wiley/AWS put it in 2025: *"Innovation is slowing down because researchers can't access the knowledge that already exists."* He contrasts AI agents limited to abstracts with those that could use "full experimental context" (methods, results) if full-text were integrated (www.linkedin.com). In practical terms, drug discovery models or materials science predictors trained only on abstracts may miss critical cues found in the body of papers. Conversely, if these models had rich text, they *could* generate hypotheses or flag connections humans overlooked.

Shift to "Agent-Ready" Ecosystems. The industry is moving toward content infrastructures designed for AI. AWS and others now offer "Agent Knowledge Bases" for scientific fields (www.linkedin.com). These services bundle peer-reviewed content (via publisher agreements) into machine-readable JSON/API form. They deliver either "complete full-text" or structured sections. By supporting the Model Context Protocol, these systems enable AI developers to *"integrate scientific content seamlessly into their existing development environments without wrestling with outdated data formats"* (www.linkedin.com). This is a promising trend: instead of hacking PDFs, AI tools will query APIs that return just the needed excerpt. But adoption is nascent, and who pays (users vs publishers) remains in flux.

Knowledge Equity. Looking ahead, if AI research assistants become ubiquitous, the gap between well-funded labs and under-resourced ones could widen. Prestigious institutions often subscribe to most major journals; lower-tier institutions may not. If their students rely on AI bots that in turn rely on those subscriptions, the latter group will effectively see a censored knowledge base. Wingfield & Wingfield warn that without corrective action, *"AI-assisted scholarship may become a self-reinforcing echo chamber"*, where only research from liberal-access journals (or those paying APCs) is amplified (www.scielo.org.za). To counter this, some suggest national or international mandates: for example, the South African policy roadmap recommends all scholarly publishing funders require AIs to be disclosed in literature synthesis, and libraries to negotiate "AI-readable licenses" as part of big deals (www.scielo.org.za).

The Future of Open Access. On the positive side, both policy and culture are shifting. The Gates Foundation's refreshed OA pledge includes *"require grantees to share preprints"* (www.gatesfoundation.org). If every funder did similarly, the corpus of accessible pre-publication manuscripts could skyrocket, giving AI far more text to chew on. Preprint servers demonstrated during COVID-19 that rapid, open sharing is possible at scale. Movements like Plan S (forcing immediate OA) and agreements to waive APCs in middle-income countries will gradually reduce the locked-out proportion.

However, economics remain tricky. Publishers rely on subscription and APC revenues, and may resist losing that…unless forced. In the short run, they may seek middle-ground: tiered APIs, micropayments for AI usage, or pay-per-crawl schemas. Industry content providers (provider like Perlego for textbooks) are starting to think about *"pay-per-search"* or *"AI credits"* models. The legal landscape might also change: if enough AI companies face copyright lawsuits (as is happening for general-purpose LLMs), courts may clarify what constitutes allowable content use. If US law institutes a broad TDM exception (as many countries have), it could dramatically ease access for academic AI development.

Conclusion

This report argues that full-text access deserves priority in the nascent field of AI-powered research assistance. Unlike generic web search, academic literature has multiple layers of gatekeeping (paywalls, publisher policies, format heterogeneity). The evidence shows that **without systemic change, AI tools will perpetuate knowledge gaps** (www.scielo.org.za) (www.gatesfoundation.org). To build truly effective and fair literary AI, we must solve the content problem. This will require coordinated efforts: continued expansion of open-access mandates, innovative licensure schemes for AI, and technical standards (like MCP) that let machines respect entitlements while retrieving full-text.

The stakes are high. As the Gates Foundation notes, *"When researchers can see what others have learned, they can build on it… That's how scientific advances are made."* (www.gatesfoundation.org) This report therefore calls for treating full-text access as a priority for accelerating science. The central concern of this report is ensuring that AI tools can *read the actual research*. We must bridge that gap with policy, partnership, and innovation, or else these AI assistants will remain only as good as the scraps of text we feed them (www.scielo.org.za).

Table 1. *Comparison of AI research tools by content access. A "✓" indicates capability or source. "Open OA" means content already free (e.g. arXiv); "Subscr. access" means content requiring a paid license or user credentials*.

Tool / Service	Coverage (No. of Papers)	Primary Content Sources	Has Paywalled/Journals?
Semantic Scholar	~175M papers (2019) (docslib.org)	Publisher metadata, abstracts Open corpora via Unpaywall (docslib.org)	✗ (does not itself host paywalled text)
Elicit	Systematic Review workflows	Full-text screening and extraction, with browser-extension retrieval on Pro or higher plans (Elicit documentation)	Authorized retrieval uses active institutional browser sessions and the institution's subscriptions.
Consensus.app	Full-text availability varies by paper	Publisher partnerships and open-access PDFs for in-platform analysis (Consensus documentation)	Participating paywalled articles can be analyzed; viewing or downloading the article depends on the user's subscription.

Tool / Service	Coverage (No. of Papers)	Primary Content Sources	Has Paywalled/Journals?
Scite.ai	Article access via MCP	Scite MCP's access resolver checks user entitlements (Scite MCP documentation)	Access routing depends on user entitlements.
ChatGPT (Deep Res.)	Sources depend on configuration	Public web, relevant files and connected app context; Consensus supports Deep Research (official OpenAI documentation; Consensus changelog)	Additional scientific content can come through connected services; publisher access remains subject to entitlements (Scite MCP documentation).
Google NotebookLM	Web plus uploaded docs	Web search & user files	✗ (same limitation as ChatGPT)
Connected Papers	60M+ (via Semantic)	Structured citations (SemScholar DB)	✗ (no full-text retrieval)
SciSummary AI	N/A (proprietary)	Likely Crossref + open archives	✗ (focus on open summaries)

Sources: Tool documentation and reviews (tamu.libguides.com) (techpoint.africa) (scite.ai).

In sum, the **full-text barrier** combines legal, economic, and technical hurdles. AI researchers and tool-builders must navigate subscriptions, copyright, and formats to get the raw material they need. Although progress is being made via APIs and policy reforms, until the majority of scholarship is genuinely open, creating truly comprehensive AI literature tools will remain extraordinarily difficult. This report therefore calls for making full-text access a priority in the new AI-driven research paradigm.

References: All claims above are supported by the cited literature and industry reports, including studies on access models (pmc.ncbi.nlm.nih.gov) (docslib.org), analyses of AI system behavior (techpoint.africa) (www.scielo.org.za), and announcements from publishers and labs (scite.ai) (www.getfulltextresearch.com).

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2025.1544965/full
2. https://www.copyright.com/blog/abstracts-vs-full-text-literature-text-mining
3. https://techpoint.africa/guide/chatgpt-deep-research?_rt=MXwxfHtzZWFyY2hfdGVybV9zdHJpbmd9fDE3NjEyNjI0NDU&_rt_nonce=c3db4dd422
4. https://www.authorsalliance.org/2024/12/06/restricting-innovation-how-publisher-contracts-undermine-scholarly-ai-research
5. https://www.scielo.org.za/scielo.php?pid=S0038-23532025000600014&script=sci_arttext&tlng=en
6. https://scite.ai
7. https://tech.yahoo.com/ai/articles/openai-launches-free-prism-app-120500641.html
8. https://www.gatesfoundation.org/ideas/articles/research-paywall-open-access
9. https://pmc.ncbi.nlm.nih.gov/articles/PMC11979259

10. https://learn.chatgpt.com/docs/web-search
11. https://help.consensus.app/en/articles/11954907-consensus-product-changelog
12. https://www.researchsolutions.com/blog/connect-your-ai-tools-to-scientific-research-with-the-scite-mcp
13. https://www.tomsguide.com/ai/ai-chatbots-are-changing-how-we-access-paywalled-news-heres-how-that-affects-you
14. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML?uri=CELEX:32019L0790
15. https://www.elsevier.com/about/policies-and-standards/text-and-data-mining
16. https://help.consensus.app/en/articles/10055108-consensus-research-database
17. https://developers.openai.com/api/docs/models/gpt-4.1
18. https://www.linkedin.com/pulse/future-scientific-discovery-why-full-text-access-missing-dale-morgan-3oi0e
19. https://docslib.org/doc/7357567/semantic-scholars-newly-expanded-coverage-reveals-how-open
20. https://tamu.libguides.com/c.php?g=1289555
21. https://support.elicit.com/en/articles/14759113-elicit-browser-extension
22. https://www.getfulltextresearch.com/getftr-enables-ai-tools-to-check-access-rights-for-academic-content
23. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
24. https://github.com/allenai/s2orc/blob/master/README.md
25. https://scite.ai/data
26. https://www.getfulltextresearch.com/ai-tools-getftr
27. https://consensus.app/home/features/full-text
28. https://www.tomsguide.com/ai/notebooklm-can-now-browse-the-web-with-deep-research-i-put-the-new-feature-to-the-test
29. https://support.google.com/gemininotebook/answer/16215270?hl=en
30. https://scispace.com/guides/typeset-evolving-into-scispace
31. https://docslib.org/doc/11192645/sci-hub-provides-access-to-nearly-all-scholarly-literature

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.