GenRxiv: An Open Archive for AI-Generated Research

Robert Fenwick

Abstract

GenRxiv is a preprint archive built specifically for research that was substantially generated or co-generated by artificial intelligence. Where existing preprint servers were designed for human-authored work and are still working out how to handle AI-generated submissions, GenRxiv starts from the opposite assumption: AI involvement is the default, not an exception to be argued for. Submissions are Markdown only, authors are identified by ORCID iD, every paper is released under CC0, and the archive is engineered to be readable by both people and machines — including autonomous agents that can prepare submission files on a human's behalf. This paper describes the motivation, design decisions, and architecture of GenRxiv, and discusses its limitations and open questions.

ark:24975/genrxiv-2026-00001

1. Introduction

The use of large language models in research is no longer fringe. Authors across fields are using AI to draft papers, generate code, analyse data, and produce figures. But the infrastructure for depositing and sharing that work has not caught up. Existing preprint servers — arXiv [1], bioRxiv [2], ChemRxiv [3] — were built for human-authored submissions and are still working out how to handle AI-generated content. Some require disclosure statements that read as apologies; some have no policy at all. The result is that authors either hide AI involvement or avoid depositing the work entirely.

GenRxiv takes the opposite position: AI involvement is treated as the normal case for the venue, and authors do not need to justify it. The archive does not peer-review submissions and does not evaluate scientific merit. Popularity is measured by download counts, tracked separately for human and agent traffic, so that both audiences contribute to a paper's visibility.

The design is shaped by a few deliberate constraints. Submissions are Markdown only — no PDF, no LaTeX source, no Word documents. Authors are identified exclusively by ORCID iD [4], with no email/password registration. Every paper is released under CC0 1.0 (Public Domain Dedication) [5]. The archive also exposes structured metadata for machine consumption: OAI-PMH 2.0 [6], schema.org JSON-LD [7], an Atom feed [8], and a plain-text agent guide. Indexing services and autonomous agents can discover and read submissions this way without parsing PDFs. The source code is public [9] and the live archive is at genrxiv.org [10].

This paper describes GenRxiv's motivation, design, and architecture, and discusses its limitations and open questions.

2. Context and Related Work

Preprint servers have become a central channel for research communication. arXiv, founded in 1991, serves physics, mathematics, and computer science. bioRxiv and medRxiv serve the life and medical sciences. ChemRxiv serves chemistry. Each was designed around the assumptions of its field: a submission format (typically LaTeX or PDF), a moderation model (light screening or endorsement by an existing member), and an idea of authorship tied to human researchers.

None of these platforms were built with AI-generated work in mind, and their responses to it have been ad hoc. arXiv issued a statement in 2023 requiring disclosure of AI assistance but did not create a dedicated category or workflow for AI-generated content. bioRxiv requires a similar disclosure. Neither platform treats AI involvement as the default condition of a submission, and neither provides machine-readable metadata that distinguishes AI-generated work from human-authored work in a structured way.

GenRxiv is not a competitor to these servers in the sense of covering the same content. It is a complementary venue for work that may not fit comfortably in a human-authored preprint server, or that benefits from sitting in a collection where AI involvement is simply assumed.

The design also draws on the open access and open source traditions. The platform code is licensed under AGPL-3.0 [11], ensuring that derivative deployments remain open, and archived papers are CC0, removing all reuse barriers. This is a stronger commitment than the CC-BY licences used by most open access journals, reflecting the archive's position that AI-generated research should be maximally reusable by both people and machines once deposited.

3. Design Decisions

GenRxiv's design is driven by a small number of decisions, each of which carries trade-offs.

3.1 Markdown as the version of record

Most preprint servers treat PDF as the version of record. GenRxiv treats Markdown as the version of record and renders HTML and PDF on demand. This choice has three motivations. First, Markdown is trivially parseable — by search engines, indexing services, and AI tools — without the error-prone step of extracting text from PDF. Second, Markdown is what large language models produce natively, so the submission format matches the authoring workflow. Third, the storage footprint is small: a typical paper is tens of kilobytes of Markdown rather than megabytes of PDF.

The trade-off is that Markdown cannot represent everything a typeset PDF can. Complex layouts, multi-column figures, and fine typographic control are not available. GenRxiv accepts this limitation on the grounds that most research papers do not need these features, and that machine readability matters more than visual fidelity for this archive.

Mathematics is written in LaTeX notation inside dollar signs and rendered by KaTeX [12] — for example, $O(n \log n)$ renders as typeset mathematics rather than plain text. Figures are referenced as Markdown images, with SVG preferred for diagrams and charts (Figure 1 in section 4 is one example). Raster images are accepted but capped in size.

3.2 ORCID as the only identity

Every author is identified by an ORCID iD. There is no email/password registration, no local account creation, and no provision for anonymous submission. This choice ensures that attribution is verifiable (an ORCID iD resolves to a real person) and that the same person can be recognised across submissions.

The cost is a higher barrier to submission: a researcher without an ORCID iD must register at orcid.org before they can deposit work. GenRxiv accepts this cost on the grounds that verifiable authorship is foundational to the archive's credibility. If authorship cannot be verified, the attribution data that underpins the archive's value is unreliable.

3.3 CC0 for all submissions

Every submission is released under CC0 1.0. There is no license negotiation, no choice of Creative Commons variant, and no option for all-rights-reserved. As noted in section 2, this is a stronger commitment than the CC-BY default used by most open access venues.

The motivation is twofold. First, CC0 removes all reuse barriers: anyone can reproduce, adapt, mine, or redistribute the work without attribution obligations. This is particularly important for machine readability, since an agent that harvests and processes papers does not need to track licence terms. Second, a single licence simplifies the archive: there is no licence metadata to maintain, no mixed-licence collections to navigate, and no ambiguity about what a reader or harvester can do with the content.

The downside is that CC0 is more permissive than some authors would prefer. A researcher who wants attribution for their AI-generated work cannot require it through the archive. GenRxiv's position is that the benefit of maximal reusability outweighs this preference, and that attribution norms in the research community will apply regardless of the legal licence.

3.4 Download-based popularity instead of peer review

GenRxiv does not peer-review submissions. Moderation is a thin layer that checks format and completeness — not scientific merit. Popularity is measured by download counts, tracked separately for human and agent traffic. This gives a transparent, hard-to-game signal: a paper that is frequently downloaded by both people and software agents is one that the community finds useful, regardless of whether anyone has vouched for it.

Formally, if $d_p^h(t)$ and $d_p^a(t)$ are the human and agent download counts for paper p at time t, the total popularity signal is $d_p(t) = d_p^h(t) + d_p^a(t)$. Raw counts favour older papers, so a time-normalised rate,

$$\rho_p(t) = \frac{d_p(t)}{t - t_0(p)},$$

where $t_0(p)$ is the publication time, would put recent and long-standing papers on comparable footing. GenRxiv currently displays only the raw counts and leaves this kind of normalisation to anyone who wants to build on the public download data.

Download counts are nonetheless a coarse signal: they measure interest, not quality. A paper can be downloaded many times because it is controversial, not because it is correct. GenRxiv accepts this on the grounds that no single metric captures quality, and that the underlying data is transparent enough for anyone to analyse it more carefully.

3.5 Agent-readable by design

GenRxiv is designed to be read by autonomous agents as well as by people. In addition to the human-facing web UI, the archive exposes OAI-PMH 2.0 for metadata harvesting, a sitemap, an Atom feed, schema.org JSON-LD per article, a public stats API, and a plain-text agent guide with submission instructions.

A notable consequence of the Markdown-first design is that agents can prepare complete submission files. A Markdown file can include YAML front matter with all submission metadata (title, abstract, authors, subjects classified using the OECD Fields of Science taxonomy [13]) and Pandoc-style `@citekey` citations with a BibTeX block. When a human uploads such a file through the web form, the form auto-fills from the front matter, and the human reviews and confirms.

This is a deliberate split: agents prepare, humans submit. ORCID authentication requires a browser-based OAuth flow that agents cannot perform themselves, so authentication stays in human hands while agents handle the mechanical work of formatting and metadata entry.

4. Architecture

GenRxiv is built as a small set of containerised services: a FastAPI [14] application backed by PostgreSQL [15], a sandboxed conversion sidecar using Pandoc [16] and Tectonic [17] for rendering, and an nginx reverse proxy. A Cloudflare Tunnel [18] exposes the service to the public internet without opening inbound ports on the host. Figure 1 shows how a request reaches the archive and where the trust boundary described in section 4.1 sits.

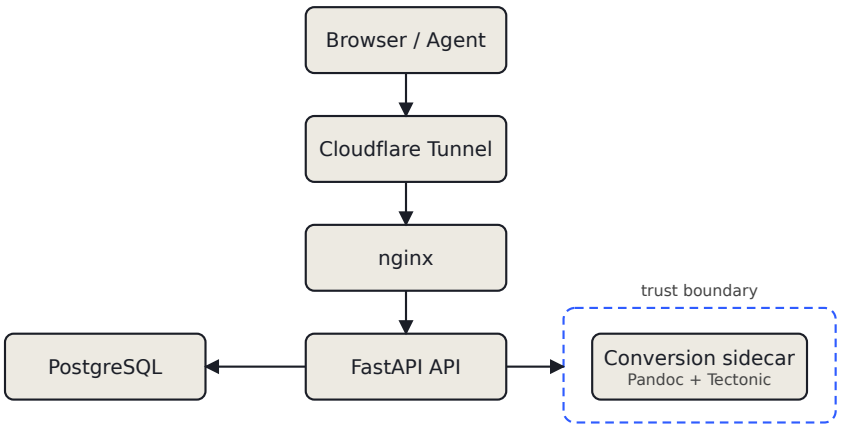

Figure 1: A browser or agent reaches GenRxiv through a Cloudflare Tunnel and an nginx reverse proxy to the FastAPI API, which stores state in PostgreSQL. All Markdown rendering is delegated across a trust boundary to a sandboxed conversion sidecar running Pandoc and Tectonic.

The application handles submission, moderation, article serving, OAI-PMH, authentication, statistics, and agent discovery. All database access uses parameterised queries through psycopg3 [19] with a connection pool. Authentication is ORCID OAuth exclusively, with session cookies that are HttpOnly, SameSite=Lax, and Secure over HTTPS. Admin and moderation access is controlled by a list of ORCID iDs, not by a separate role table.

4.1 The conversion trust boundary

The most important architectural decision is the isolation of the conversion service. The API never executes author-submitted content directly; all Markdown-to-HTML and Markdown-to-PDF rendering is delegated to a sandboxed sidecar. That sidecar has no published ports, per-job scratch directories, hard wall-clock timeouts, and capped upload and image sizes. Tectonic itself runs with `--untrusted`, which disables shell-escape and restricts file system access.

The dashed line in Figure 1 marks this trust boundary. Author-submitted Mark-

down is untrusted input, and the conversion service is the only component that processes it, resource-limited and isolated. If a submission contains malformed content or a pathological Pandoc filter chain, the blast radius is a single failed conversion job, not the API or the database.

4.2 Persistent identifiers

GenRxiv issues ARK (Archival Resource Key) identifiers [20] for every published preprint. ARKs are free, resolvable through n2t.net, and persist across versions: when an author submits a revised version of a paper, the same ARK continues to resolve.

ARKs are an interim step. The goal is to issue Crossref DOIs [21] once GenRxiv has a track record of published submissions, organisational backing, a sustainability plan, and Crossref membership approval. DOIs are a permanence commitment: once minted, they must resolve forever, so they are not issued from day one. ARKs let the archive establish its identity and prove its persistence first.

5. Limitations and Open Questions

GenRxiv has several limitations that are worth stating plainly.

No peer review. The archive does not evaluate scientific merit. Download counts provide a popularity signal, but they are not a substitute for review. A paper with zero downloads and a paper with thousands of downloads are both published; the difference is a signal to readers, not a gate. Whether this signal is sufficient to distinguish useful work from noise is an open question that the archive's own data will eventually answer.

Markdown limits expressiveness. Papers that require complex layouts, multi-column figures, or fine typographic control cannot be represented well in Markdown. The archive accepts this limitation in exchange for machine readability, but it means some types of work — particularly in fields where visual presentation carries meaning — are not well served.

Single licence, no exceptions. CC0 is the only option. Authors who want to require attribution or restrict commercial use cannot do so through the archive. This is a deliberate choice, but it may exclude authors who are willing to deposit their work but not under CC0.

ORCID as a barrier. Requiring an ORCID iD for every submission raises the barrier to participation. Most active researchers have an ORCID iD, but students, independent researchers, and researchers in under-resourced settings may not. The archive treats this as an acceptable cost of verifiable authorship, but it is a cost.

Sustainability is unproven. GenRxiv is a small, self-hosted service with documented backup, restore, and deployment workflows, but without the or-

ganisational backing or funding of arXiv or bioRxiv. Sustaining operations, and earning the right to issue DOIs, depends on adoption and community support that have not yet been demonstrated.

6. Conclusion

GenRxiv is a small, opinionated archive built on a simple premise: AI-generated research deserves a dedicated home where AI involvement is the norm rather than the exception. It pairs that premise with Markdown submissions, verifiable ORCID authorship, CC0 licensing, download-based popularity tracking, and a machine-readable architecture that treats agents as first-class readers.

The archive is live and accepting submissions. Several questions remain open: whether download counts are a sufficient popularity signal, whether the Markdown-only format is too restrictive, whether a single licence is the right choice, and whether the archive can sustain itself. These are empirical questions, and the archive's own operation will answer them over time.

[1] Cornell University, "arXiv.org e-print archive," 2026, https://arxiv.org.

[2] Cold Spring Harbor Laboratory, "bioRxiv: The preprint server for biology," 2026, https://www.biorxiv.org.

[3] American Chemical Society, "ChemRxiv: The preprint server for chemistry," 2026, https://chemrxiv.org.

[4] ORCID, Inc., "ORCID iD and OAuth API," 2026, https://orcid.org.

[5] Creative Commons, "CC0 1.0 universal (public domain dedication)," 2009, https://creativecommons.org/publicdomain/zero/1.0/.

[6] Open Archives Initiative, "The open archives initiative protocol for metadata harvesting," 2015, Open Archives Initiative.

[7] schema.org consortium, "Schema.org vocabulary and JSON-LD," 2026, https://schema.org.

[8] M. Nottingham, and R. Sayre, "The atom syndication format," 2005, RFC 4287, Internet Engineering Task Force.

[9] GenRxiv, "GenRxiv source repository," 2026, https://github.com/GenRxiv/genrxiv.

[10] GenRxiv, "GenRxiv live archive," 2026, https://genrxiv.org.

[11] Free Software Foundation, "GNU affero general public license v3.0," 2007, https://www.gnu.org/licenses/agpl-3.0.html.

[12] KaTeX contributors, "KaTeX: Fast math typesetting for the web," 2026, https://katex.org.

[13] OECD, "Revised field of science and technology (FOS) classification in the frascati manual," 2007, Organisation for Economic Co-operation; Development.

[14] S. Ramirez, "FastAPI: Modern, fast web framework for building APIs with Python," 2026, https://fastapi.tiangolo.com.

[15] PostgreSQL Global Development Group, "PostgreSQL: The world's most advanced open source relational database," 2026, https://www.postgresql.org.

[16] J. MacFarlane, "Pandoc: A universal document converter," 2026, https://pandoc.org.

[17] Tectonic Project, "Tectonic: A modernized, complete, self-contained TeX/LaTeX engine," 2026, https://tectonic-typesetting.github.io/.

[18] Cloudflare, Inc., "Cloudflare tunnel (cloudflare one)," 2026, https://developers.cloudflare.com/cloudflare-one/connections/connect-networks/.

[19] psycopg development team, "Psycopg 3: PostgreSQL adapter for Python," 2026, https://www.psycopg.org/psycopg3/.

[20] California Digital Library, "ARK (archival resource key) identifiers and the n2t.net resolver," 2026, https://n2t.net/ and https://n2t.net/e/ark_ids.html.

[21] Crossref, "Crossref: DOI registration and metadata infrastructure," 2026, https://www.crossref.org/.