Towards LLM-Powered Automation of a Dark Matter Constraint Repository

Lanqing Yuan
Department of Physics
Washington University in St. Louis
yuan.l@wustl.edu

Karthik Ramanathan
Department of Physics
Washington University in St. Louis
karthikr@wustl.edu

Abstract

Dark matter constraint repositories are critical community infrastructure, giving experimentalists and theorists a shared landscape of existing bounds. Yet the most widely-used repositories are maintained by individual volunteers, creating a sustainability risk as the pace of new results accelerates. We present a large language model (LLM) pipeline that monitors arXiv, extracts limit curves from papers, integrates them as code, and opens pull requests (PRs) for human review. On a 346-paper benchmark whose ground truth is the upstream-curated repository itself, the pipeline classifies the coupling type correctly for 90.5% of papers and reaches a median coupling residual of 0.33 dex (a factor of two for 48% of curves), with 76% mean mass-range coverage. This is driven by treating each extraction as a noisy sample reconciled through consensus voting, a physics convention-canonicalization layer built with the agentic physics assistant Get Physics Done (GPD), and a scoring methodology that separates genuine extraction error from non-comparability. The remaining difficulty is concentrated in rare coupling types with idiosyncratic conventions (macro-averaged residual 1.1 dex). The pipeline is deployed and has generated limit proposals; none have merged — governance of AI-generated scientific data is itself an unsolved problem.

1 Introduction

The experimental search for ultralight dark matter is undergoing a "renaissance" [1], with new proposals targeting masses from 10^{-22} eV to several eV [2, 3]. The result is an ever-denser landscape of constraints across at least 15 distinct *coupling types* (ways dark matter can interact with the Standard Model, e.g. axion–photon or dark photon kinetic mixing) [4], each with its own units, parameterisations, and confidence levels. Dark matter constraint repositories compile these bounds into publication-quality *exclusion plots* — the mass–coupling regions ruled out — used across the field [5, 4, 6]. The AxionLimits repository [5] tracks over 600 limits and 160 projected sensitivities, yet is maintained by a single volunteer who must find new results, digitise figures or tables, convert parameterisations (e.g. χ vs. g [4]), write plotting code, and regenerate figures.

We present a pipeline that partially automates this workflow: (1) **Discovery** via daily arXiv monitoring with keyword filtering and LLM relevance classification; (2) **Extraction** using a two-stage approach (text/tables first, vision fallback for plot-only papers) with multi-read consensus voting and a physics convention-canonicalization layer; (3) **Integration** through automated code generation with Abstract Syntax Tree (AST)-based insertion, notebook updates, and plot regeneration; and (4) **Review** via PRs

*This paper was drafted with the assistance of large language models (Claude) under human supervision and editorial control.

Conference on Physics and AI at Stanford University (PAI 2026).

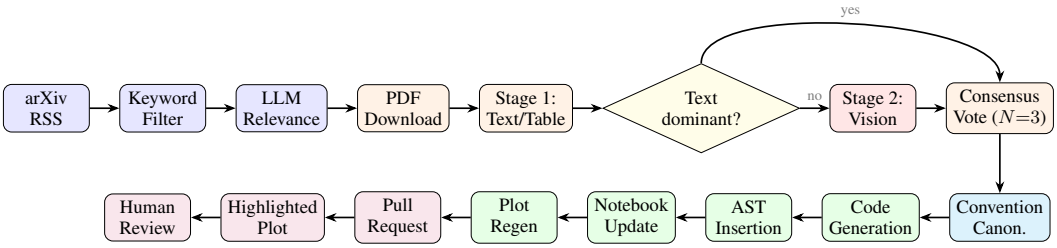

Figure 1: Pipeline architecture. Papers flow from arXiv through keyword and LLM filtering, two-stage extraction (text-first, vision fallback), multi-read consensus voting, physics convention canonicalization, code generation with AST-based insertion, plot regeneration, and finally a pull request with a highlighted constraint plot for human review.

with highlighted constraint plots. The system also includes a weekly preprint version checker and a historical backfill workflow using INSPIRE-HEP [7].[1]

Contributions. To our knowledge, no existing system automates the full loop from discovery through extraction, code integration, and human review for a live scientific repository. (i) An end-to-end pipeline that turns an arXiv paper into a reviewable code change for a real constraint repository. (ii) An extraction stack that treats LLM reads as noisy samples and reconciles them by consensus with source-quality tiering, and pairs them with a physics convention-canonicalization layer whose conversions were established with the agentic physics assistant Get Physics Done (GPD) [9] rather than by prompting. (iii) A 346-paper benchmark and scoring methodology that grounds truth in the upstream repository and separates extraction error from non-comparability. (iv) The empirical finding that, with these in place, log–log plot reading is no longer the dominant error source — the frontier shifts to deterministic, convention-bound rare coupling types.

2 System Architecture

2.1 Two-Stage Extraction with Consensus Voting

Figure 1 shows the end-to-end pipeline. The extraction pipeline prioritises text and table data (Stage 1) over vision-based figure reading (Stage 2), motivated by both cost and accuracy.

Stage 1: Text and tables. The PDF is parsed with PyMuPDF [10] and the full text sent to an LLM [8] with a structured prompt requesting JSON output: coupling type, mass–coupling data points, the model's declared units/convention, and metadata. Prompt-injection defences strip control characters and wrap content in sentinel delimiters.

Stage 2: Vision fallback. Vision is invoked only when Stage 1 is not *clearly dominant* (a valid mass window, ≥ 5 points, sufficient confidence, non-degenerate). Selected figure pages are rendered to images; the model reads axis labels, tick marks, and curve positions from log–log exclusion plots.

Consensus across noisy reads. LLM extraction is non-deterministic, so each paper is read several times independently (N=3) and reconciled. The coupling type is chosen by majority vote; the reported curve is the *medoid* — the candidate with the smallest median pairwise distance to the others in log–log space. The medoid is a real sample, not an average, so consensus rejects one-off mis-reads without ever fabricating a non-physical curve by averaging incompatible ones.

Quality-tier source selection. When candidates disagree, they are ranked by a lexicographic quality tuple rather than by raw point count: validity (values in physically allowed ranges), non-degeneracy (a traced curve must span ≥ 1 dex, not a flat artefact), recoverability, *source tier* (table $>$ text $>$ figure-vision), corroboration between reads, and confidence. A sparse text or table snippet (≤ 3 points, e.g. a headline bound quoted in prose) is demoted below a traceable multi-point figure curve, so a single quoted number cannot override an actual digitised constraint.

[1] Code, evaluation benchmark, and prompts: `https://github.com/FaroutYLq/AutoAxionLimits`

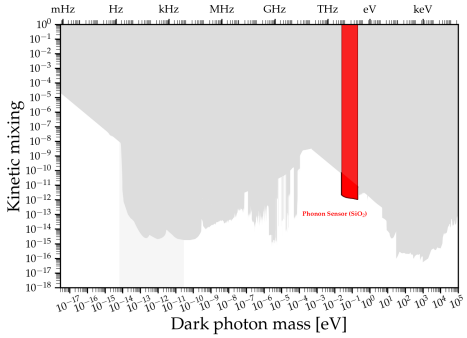

Figure 2: Highlighted dark photon constraint projection plot from a production pull request for a phonon-sensor proposal [12]. Existing limits in grey; new proposal in red. This format lets reviewers verify the proposal visually.

2.2 Convention Canonicalization

The same physics is reported in incompatible y-variables across papers and even across files within one coupling type, and the conversion often lives in the repository's plotting code, not the paper. For example, axion–nucleon couplings are stored as the derivative coupling $g_{aNN} = C_N/2f_a$ [GeV^{-1}] (with model coupling C_N, axion decay constant f_a, and nucleon mass m_N) but plotted dimensionless after an in-code $\times 2m_N$ rescaling — with a per-file $\times m_N$ exception for one detector. Other families hide analogous factors (dark-photon ε vs. ε^2, scalar d_e vs. fifth-force $\sqrt{\alpha}$) or use large sentinel values that are plotting artefacts, not data.

These *deterministic* mismatches are unfixable by prompting, since the conversion depends on per-file repository storage absent from the paper. We encode them in a per-file registry that canonicalizes both extracted and reference curves before scoring, flagging unmappable cases `[CONVENTION REVIEW]` rather than emitting a silent wrong number. The conversion factors were not guessed but derived with GPD, *code-verified* against the plotting source and *citation-audited* against the literature — a small but exacting physics task whose rigour is what makes the registry trustworthy.

2.3 Code Generation, Integration, and Review

The LLM generates a `@staticmethod` for the relevant plotting class (a post-generation guard enforces the decorator); insertion uses Python's AST module to locate the last method in the target class — never regex — preventing corruption. Notebook cells are inserted via `nbformat` and plots regenerated headlessly via `nbconvert`. Every update becomes a pull request, with low-confidence or unresolved-convention extractions flagged in the title. Crucially, we render a "highlighted" constraint plot in which all existing limits are greyed out and only the new proposed limit is shown in red (Figure 2), enabling rapid visual verification by domain experts without reading code.

2.4 Workflows and Cost

The **daily arXiv digest** queries, filters, classifies relevance, and processes matching papers. The **weekly preprint checker** scans existing data files for arXiv IDs and re-extracts on version updates to catch changed results; one PR detected that a JWST dark photon search [11] had changed a limit to a projection upon publication. The **historical backfill** searches INSPIRE-HEP [7] for older papers by citation count, splitting large jobs across daily runs. The pipeline runs on a frontier model (Claude Opus [8]), at a per-paper cost ranging from a few cents for text-only papers to a few dollars for those requiring the vision fallback, depending on the number of model calls and image tokens.

3 Evaluation and Discussion

Benchmark and ground truth. We evaluate on 346 papers spanning 14 coupling types, selected as all curated data files [5] with a known arXiv ID. The ground truth (GT) for each is the upstream repository's own digitised curve — never data digitised by our own model — which avoids circularity

Table 1: Benchmark performance on 346 papers (271 comparable, 243 with mass-range overlap; 0.3 dex $\approx$ a factor of two). (b) gives the same median residual and within-factor-2 rate over subsets: by winning source; by pooling (*micro* = over papers, so abundant types dominate; *macro* = over the per-type medians, weighting every coupling type equally); and for the best/worst types.

(a) Aggregate metrics

Metric	Value
Coupling-type accuracy	90.5%
Median residual	0.331 dex
Within factor 2	48.4%
Within factor 3	61.7%
Mean interp. coverage	76.2%
Zero-overlap papers	10.3%

(b) Residual by subset

Subset	Med. resid.	$\leq$0.3 dex
Text (N=148)	0.326 dex	49.7%
Figure-vision (N=119)	0.338 dex	46.8%
Micro-avg (per paper)	0.331 dex	48.4%
Macro-avg (per type)	1.115 dex	31.1%
AxionPhoton (N=127)	0.207 dex	60.9%
AxionEDM (N=3)	4.649 dex	1.7%

but carries a consequence: the reference was itself digitised and rescaled from the same papers, so a *perfect* extraction still shows a nonzero residual equal to the upstream digitisation/convention gap. Our reported residuals are therefore an *upper bound* on true extraction error.

Metric. We compare in $\log_{10}$–$\log_{10}$ space: a log–log interpolation of the extracted points is evaluated at the ground-truth masses, giving a per-point residual $|\log_{10} g_{\text{ext}} - \log_{10} g_{\text{truth}}|$ (0.3 dex $\approx$ factor of 2). We report median residual and *interpolation coverage* (fraction of GT masses the curve spans) separately, since a curve can be accurate where it overlaps yet cover the wrong mass window. A curve is scored only against a same-coupling GT curve; non-comparable cases (wrong coupling, incompatible convention, single-point or unusable GT, no extracted points) are tracked separately rather than polluting the residuals.

Results. Coupling-type classification (against the repository's directory structure) is reliable (90.5%), and the property labels (*is new limit*, *is projection*, *data source*), graded against an independent LLM labeler, match a human audit 15/15, 15/15, and 14/15. On extraction, 90% of comparable papers achieve mass-range overlap, with a median residual of 0.33 dex, 48% of curves within a factor of two, and 76% mean coverage (Table 1). The most consequential result is the *source breakdown*: figure-vision extraction (0.338 dex) is now on par with text (0.326 dex). Reading dense log–log plots — historically the hard part of this task — is no longer the limiting factor; consensus voting and quality-tier selection over multiple reads close most of the figure-vision gap.

Where the difficulty now lives. The frontier has shifted to *rare coupling types*: the per-paper micro-average (0.33 dex) is dominated by AxionPhoton (N=127, 0.21 dex), but the per-type macro-average is 1.1 dex, as small-sample, convention-heavy types lag (AxionNeutron 0.77, VectorBL 0.91, AxionEDM 4.6 dex). AxionEDM is the clearest case — its operator coupling is flat in mass while the reported *response* $\propto 1/m_a$ folds in the field amplitude, a convention our registry flags but does not yet auto-convert — and these types are also misclassified most often, dropping them from the comparable set entirely and compounding their difficulty.

The worst errors are deterministic physics, not perception. The largest residuals were never about pixels but unit/convention mismatches — the axion–nucleon $2m_N$ factor, the scalar d_e vs. $\sqrt{\alpha}$ 10–15 dex gap, dark-photon ε vs. ε^2, sentinel values mistaken for data — that are fixed, not stochastic, and depend on per-file repository conventions absent from the paper. No prompt fixes them; the high-leverage fix was a separate, auditable physics layer. We used GPD to produce a *code-verified, citation-audited* per-file convention registry, with a `[CONVENTION REVIEW]` flag for anything outside the verified set. The general lesson: automated scientific extraction couples a *stochastic* perception problem (best handled by sampling and voting) with a *deterministic* domain-knowledge problem (best handled by verified derivation), and the two demand different tools.

Open problems and governance. Rare coupling types remain hard, and at 90% coupling accuracy one proposal in ten targets the wrong plot, so human review stays essential. The pipeline is deployed but no proposal has merged, as no review process for AI-generated scientific data exists: we recommend a community editorial board akin to PDG reviewers [13], and publishing machine-readable limit data alongside papers would bypass the extraction bottleneck entirely.

References

[1] C. B. Adams et al., "Axion dark matter," in *Snowmass 2021*, arXiv:2203.14923, 2022.

[2] I. G. Irastorza and J. Redondo, "New experimental approaches in the search for axion-like particles," *Prog. Part. Nucl. Phys.* **102**, 89, 2018.

[3] J. E. Kim and G. Carosi, "Axions and the strong CP problem," *Rev. Mod. Phys.* **82**, 557, 2010.

[4] A. Caputo, A. J. Millar, C. A. J. O'Hare, and E. Vitagliano, "Dark photon limits: a handbook," *Phys. Rev. D* **104** (9), 095029, 2021, arXiv:2105.04565 [hep-ph].

[5] C. A. J. O'Hare, "AxionLimits," `https://github.com/cajohare/AxionLimits`, 2020.

[6] C. A. J. O'Hare, "Cosmology of axion dark matter," *PoS* **COSMICWISPers**, 040, 2024.

[7] INSPIRE Collaboration, "INSPIRE-HEP: High-Energy Physics Literature Database," `https://inspirehep.net`, 2024.

[8] Anthropic, "The Claude model family," `https://www.anthropic.com`, 2024.

[9] Physical Superintelligence PBC, "Get Physics Done (GPD)," v1.2.2, 2026. `https://github.com/psi-oss/get-physics-done`.

[10] Artifex Software, "PyMuPDF," `https://github.com/pymupdf/PyMuPDF`, 2015–2025.

[11] H. An, S. Ge, J. Liu, and Z. Lu, "Direct Detection of Dark Photon Dark Matter with the James Webb Space Telescope," arXiv:2402.17140, 2024.

[12] I. M. Bloch, S. Knapen, X. Li, A. Madden, and G. Marocco, "Dark Matter Detection Using Phonon Sensing in Amorphous Materials," arXiv:2603.22390, 2026.

[13] S. Navas et al. (Particle Data Group), "Review of particle physics," *Phys. Rev. D* **110** (3), 030001, 2024.